Quantifying and Mitigating the Impact of Label Errors on Model Disparity Metrics

Julius Adebayo,Melissa Hall,Bowen Yu,Bobbie Chern

Errors in labels obtained via human annotation adversely affect a trained model's performance. Existing approaches propose ways to mitigate the effect of label error on a model's downstream accuracy, yet little is known about its impact on a model's group-based disparity metrics\footnote{Group-based disparity metrics like subgroup calibration, false positive rate, false negative rate, equalized odds, and equal opportunity are more often known, colloquially, as \textit{fairness metrics} in the literature. We use the term group-based disparity metrics in this work.}. Here we study the effect of label error on a model's group-based disparity metrics like group calibration. We empirically characterize how varying levels of label error, in both training and test data, affect these disparity metrics. We find that group calibration and other metrics are sensitive to train-time and test-time label error---particularly for minority groups. For the same level of label error, the percentage change in group calibration error for the minority group is on average 1.5 times larger than the change for the majority group. Towards mitigating the impact of training-time label error, we present an approach to estimate how changing a single training input's label affects a model's group disparity metric on a test set. We empirically assess the proposed approach on a variety of datasets and find a 10-40\% improvement, compared to alternative approaches, in identifying training inputs that improve a model's disparity metric. The proposed approach can help surface training inputs that may need to be corrected for improving a model's group-based disparity metrics.

****

Suppression helps: Lateral Inhibition-inspired Convolutional Neural Network for Image Classification

Chengyuan Zhuang,Xiaohui Yuan,XUAN GUO

Convolutional neural networks (CNNs) have become powerful and popular tools since deep learning emerged for image classification in the computer vision field. For better recognition, the dimensions of depth and width have been explored, leading to convolutional neural networks with more layers and more channels. In addition to these factors, neurobiology also suggests the widely existing lateral inhibition (e.g., Mach band effect), which increases the contrast of nearby neuron excitation in the lateral direction, to help recognition. However, such an important mechanism has not been well explored in modern convolutional neural networks. In this paper, we explicitly explore the filter dimension in the lateral direction and propose our lateral inhibition-inspired (LI) design. Our naive design incorporates the low-pass filter, while eliminating the central weight to mimic the inhibition strength decay. The inhibition value is computed from the filtering result of the input, with a simple learnable weight parameter per channel for multiplication to decide the strength. Then the inhibition value is subtracted from the input as suppression, which could increase the contrast to help recognition. We also suggest an alternative using depthwise convolution, as a general form. Our design could work on both the plain convolution and the convolutional block with residual connection, while being compatible with existing modules. Without any channel attention along the channel dimension, the preliminary results demonstrate an absolute improvement of 3.68\% and 0.69\% over AlexNet and ResNet-18, respectively, in the ImageNet data set, with little increase in parameters, indicating the merits of our design to help feature learning for image classification.

****

Factorized Fourier Neural Operators

Alasdair Tran,Alexander Mathews,Lexing Xie,Cheng Soon Ong

We propose the Factorized Fourier Neural Operator (F-FNO), a learning-based approach for simulating partial differential equations (PDEs). Starting from a recently proposed Fourier representation of flow fields, the F-FNO bridges the performance gap between pure machine learning approaches to that of the best numerical or hybrid solvers. This is achieved with new representations – separable spectral layers and improved residual connections – and a combination of training strategies such as the Markov assumption, Gaussian noise, and cosine learning rate d ecay. On several challenging benchmark PDEs on regular grids, structured meshes, and point clouds, the F-FNO can scale to deeper networks and outperform both the FNO and the geo-FNO, reducing the error by 83% on the Navier-Stokes problem, 31% on the elasticity problem, 57% on the airfoil flow problem, and 60% on the plastic forging problem. Compared to the state-of-the-art pseudo-spectral method, the F-FNO can take a step size that is an order of magnitude larger in time and achieve an order of magnitude speedup to produce the same solution quality.

****

DFPC: Data flow driven pruning of coupled channels without data.

Tanay Narshana,Chaitanya Murti,Chiranjib Bhattacharyya

Modern, multi-branched neural network architectures often possess complex interconnections between layers, which we call coupled channels (CCs). Structured pruning of CCs in these multi-branch networks is an under-researched problem, as most existing works are typically designed for pruning single-branch models like VGG-nets. While these methods yield accurate subnetworks, the improvements in inference times when applied to multi-branch networks are comparatively modest, as these methods do not prune CCs, which we observe contribute significantly to inference time. For instance, layers with CCs as input or output take more than 66% of the inference time in ResNet-50. Moreover, pruning in the data-free regime, where data is not used for pruning, is gaining traction owing to privacy concerns and computational costs associated with fine-tuning. Motivated by this, we study the problem of pruning CCs in the data-free regime. To facilitate the development of algorithms to prune CCs, we define Data Flow Couplings (DFCs) to enumerate the layers that constitute coupled connections and the associated transformation. Additionally, saliencies for pruning CCs cannot be gauged in isolation, as there may be discrepancies among the layerwise importance of CCs using conventional scoring strategies. This necessitates finding grouped saliencies to gauge the importance of all corresponding coupled elements in a network. We thus propose the Backwards Graph-based Saliency Computation (BGSC) algorithm, a data-free method that computes saliencies by estimating an upper bound to the reconstruction error of intermediate layers; we call this pruning strategy Data Flow driven Pruning of Coupled channels (DFPC). Finally, we show the efficacy of DFPC for models trained on standard datasets. Since we pruned coupled channels, we achieve up to 1.66x improvements in inference time for ResNet-101 trained on CIFAR-10 with a 5% accuracy drop without fine-tuning. With access to the ImageNet training set, we achieve significant improvements over the data-free method and see an improvement of at least 47.1% in speedup for a 2.3% accuracy drop for ResNet-50 against our baselines.

****

TVSPrune - Pruning Non-discriminative filters via Total Variation separability of intermediate representations without fine tuning

Chaitanya Murti,Tanay Narshana,Chiranjib Bhattacharyya

Achieving structured, data-free sparsity of deep neural networks (DNNs) remains an open area of research. In this work, we address the challenge of pruning filters without access to the original training set or loss function. We propose the discriminative filters hypothesis, that well-trained models possess discriminative filters, and any non-discriminative filters can be pruned without impacting the predictive performance of the classifier. Based on this hypothesis, we propose a new paradigm for pruning neural networks: distributional pruning, wherein we only require access to the distributions that generated the original datasets. Our approach to solving the problem of formalising and quantifying the discriminating ability of filters is through the total variation (TV) distance between the class-conditional distributions of the filter outputs. We present empirical results that, using this definition of discriminability, support our hypothesis on a variety of datasets and architectures. Next, we define the LDIFF score, a heuristic to quantify the extent to which a layer possesses a mixture of discriminative and non-discriminative filters. We empirically demonstrate that the LDIFF score is indicative of the performance of random pruning for a given layer, and thereby indicates the extent to which a layer may be pruned. Our main contribution is a novel one-shot pruning algorithm, called TVSPrune, that identifies non-discriminative filters for pruning. We extend this algorithm to IterTVSPrune, wherein we iteratively apply TVSPrune, thereby enabling us to achieve greater sparsity. Last, we demonstrate the efficacy of the TVSPrune on a variety of datasets , and show that in some cases, we can prune up to 60% of parameters with only a 2% loss of accuracy without any fine-tuning of the model, beating the nearest baseline by almost 10%.

****

Finding Actual Descent Directions for Adversarial Training
Fabian Latorre,Igor Krawczuk,Leello Tadesse Dadi,Thomas Pethick,Volkan Cevher

Adversarial Training using a strong first-order adversary (PGD) is the gold standard for training Deep Neural Networks that are robust to adversarial examples. We show that, contrary to the general understanding of the method, the gradient at an optimal adversarial example may increase, rather than decrease, the adversarially robust loss. This holds independently of the learning rate. More precisely, we provide a counterexample to a corollary of Danskin's Theorem presented in the seminal paper of Madry et al. (2018) which states that a solution of the inner maximization problem can yield a descent direction for the adversarially robust loss. Based on a correct interpretation of Danskin's Theorem, we propose Danskin's Descent Direction (DDi) and we verify experimentally that it provides better directions than those obtained by a PGD adversary. Using the CIFAR10 dataset we further provide a real world example showing that our method achieves a steeper increase in robustness levels in the early stages of training, and is more stable than the PGD baseline. As a limitation, PGD training of ReLU+BatchNorm networks still performs better, but current theory is unable to explain this.

****

A Study of Biologically Plausible Neural Network: the Role and Interactions of Brain-Inspired Mechanisms in Continual Learning
Fahad Sarfraz,Elahe Arani,Bahram Zonooz

Humans excel at continually acquiring, consolidating, and retaining information from an ever-changing environment, whereas artificial neural networks (ANNs) exhibit catastrophic forgetting. There are considerable differences in the complexity of synapses, the processing of information, and the learning mechanisms in biological neural networks and their artificial counterpart, which may explain the mismatch in performance. We consider a biologically plausible framework that constitutes separate populations of exclusively excitatory and inhibitory neurons which adhere to Dale's principle and the excitatory pyramidal neurons are augmented with dendritic-like structures for context-dependent processing of stimuli. We then conduct a comprehensive study on the role and interactions of different mechanisms inspired by the brain including sparse non-overlapping representations, Hebbian learning, synaptic consolidation, and replay of past activations that accompanied the learning event. Our study suggests that employing multiple complementary mechanisms in a biologically plausible architecture, similar to the brain, can be effective in enabling continual learning in ANNs.

****

Learning Continuous Normalizing Flows For Faster Convergence To Target Distribution via Ascent Regularizations
Shuangshuang Chen,Sihao Ding,Yiannis Karayiannidis,Mårten Björkman

Normalizing flows (NFs) have been shown to be advantageous in modeling complex distributions and improving sampling efficiency for unbiased sampling. In this work, we propose a new class of continuous NFs, ascent continuous normalizing flows (ACNFs), that makes a base distribution converge faster to a target distribution. As solving such a flow is non-trivial and barely possible, we propose a practical implementation to learn flexibly parametric ACNFs via ascent regularization and apply it in two learning cases: maximum likelihood learning for density estimation and minimizing reverse KL divergence for unbiased sampling and variational inference. The learned ACNFs demonstrate faster convergence towards the target distributions, therefore, achieving better density estimations, unbiased sampling and variational approximation at lower computational costs. Furthermore, the flows show to stabilize themselves to mitigate performance deterioration and are less sensitive to the choice of training flow length T.

****

pFedKT: Personalized Federated Learning via Knowledge Transfer
Liping Yi,Xiaorong Shi,Gang Wang,xiaoguang Liu
Federated learning (FL) has been widely studied as a new paradigm to achieve multi-party collaborative modelling on decentralized data with privacy protection. Unfortunately, traditional FL suffers from Non-IID data distribution, where clients' private models after FL are even inferior to models trained standalone. Existing approaches to tackle this challenge fall into two directions: a) pursuing a better global model through mitigating biases of private models, and b) improving personalized private models by personalized federated learning (PFL). Still, both of them have limited accuracy improvements in private models. To this end, \textit{we design pFedKT, a novel personalized federated learning framework with knowledge transfer, towards boosting the performances of personalized private models on Non-IID data}. It involves two types of knowledge transfer: a) transferring \textit{historical private knowledge} to new private models by local hypernetworks; b) transferring \textit{the global model's knowledge} to private models through contrastive learning. After absorbing the historical private knowledge and the latest global knowledge, the personalization and generalization of private models are both enhanced. Besides, we derive pFedKT's generalization and prove its convergence theoretically. Extensive experiments verify that pFedKT presents $0.31\%-3.46\%$ accuracy improvements of private models than the state-of-the-art baseline.

****

FARE: Provably Fair Representation Learning
Nikola Jovanovi■,Mislav Balunovic,Dimitar Iliev Dimitrov,Martin Vechev
Fair representation learning (FRL) is a popular class of methods aiming to produce fair classifiers via data preprocessing. However, recent work has shown that prior methods achieve worse accuracy-fairness tradeoffs than originally suggested by their results. This dictates the need for FRL methods that provide provable upper bounds on unfairness of any downstream classifier, a challenge yet unsolved. In this work we address this challenge and propose Fairness with Restricted Encoders (FARE), the first FRL method with provable fairness guarantees. Our key insight is that restricting the representation space of the encoder enables us to derive suitable fairness guarantees, while allowing empirical accuracy-fairness tradeoffs comparable to prior work. FARE instantiates this idea with a tree-based encoder, a choice motivated by inherent advantages of decision trees when applied in our setting. Crucially, we develop and apply a practical statistical procedure that computes a high-confidence upper bound on the unfairness of any downstream classifier. In our experimental evaluation on several datasets and settings we demonstrate that FARE produces tight upper bounds, often comparable with empirical results of prior methods, which establishes the practical value of our approach.

****

ONLINE RESTLESS BANDITS WITH UNOBSERVED STATES
Bowen Jiang,Bo Jiang,Jian Li,TAO LIN,Xinbing Wang,Chenghu Zhou
We study the online restless bandit problem, where each arm evolves according to a Markov chain independently, and the reward of pulling an arm depends on both the current state of the corresponding Markov chain and the action. The agent (decision maker) does not know the transition kernels and reward functions, and cannot observe the states of arms all the time. The goal is to sequentially choose which arms to pull so as to maximize the expected cumulative rewards collected. In this paper, we propose TSEETC, a learning algorithm based on Thompson Sampling with Episodic Explore-Then-Commit. The algorithm proceeds in episodes of increasing length and each episode is divided into exploration and exploitation phases. In the exploration phase in each episode, action-reward samples are collected in a round-robin way and then used to update the posterior as a mixture of Dirichlet distributions. At the beginning of the exploitation phase, TSEETC gene

rates a sample from the posterior distribution as true parameters. It then follows the optimal policy for the sampled model for the rest of the episode. We establish the Bayesian regret bound $\tilde {\mathcal{O}}(\sqrt{T})$ for TSEETC, where T is the time horizon. This is the first bound that is close to the lower bound of restless bandits, especially in an unobserved state setting. We show through simulations that TSEETC outperforms existing algorithms in regret.

****

Learning to aggregate: A parameterized aggregator to debias aggregation for cross-device federated learning

Tao Shen,Kun Kuang,Yaliang Li,Feng Wang,Zheqi Lv,Hongxia Yang,Chao Wu,Fei Wu

Federated learning (FL) emerged as a novel machine learning setting that enables collaboratively training deep models on decentralized private data. Due to the heterogeneity (non-iidness) of the decentralized data, FL methods (e.g. FedAvg) suffers from unstable and slow convergence. Recent works explain the non-iid problem in FL as the client drift, and deal with it by enforcing regularization at local updates. However, these works neglect the heterogeneity among different communication rounds: the data of sampled candidates at different communication rounds are also of non-iid distribution, and we term it as period drift, which as well as client drift can lead to aggregation bias that degrade convergence. To deal with it, we propose a novel aggregation strategy, named FedPA, that uses a Parameterized Aggregator, as an alternative of averaging. We frame FedPA within a meta-learning setting, and formulates the aggregator as a meta-learner, to learn to aggregate the model parameters of clients. FedPA can directly learn the aggregation bias and well calibrate and control the direction of aggregated parameters to a better direction towards the optimum. Experiments show that FedPA can achieve competitive performances compared with conventional baselines.

****

Deep Reinforcement Learning based Insight Selection Policy

Libio Goncalves Braz,Allmin Pradhap Singh Susaiyah,Milan Petkovic,Aki Härmä

We live in the era of ubiquitous sensing and computing. More and more data is being collected and processed from devices, sensors and systems. This opens up opportunities to discover patterns from these data that could help in gaining better understanding into the source that produces them. This is useful in a wide range of domains, especially in the area of personal health, in which such knowledge could help in allowing users to comprehend their behaviour and indirectly improve their lifestyle. Insight generators are systems that identify such patterns and verbalise them in a readable text format, referred to as insights. The selection of insights is done using a scoring algorithm which aims at optimizing this process based on multiple objectives, e.g., factual correctness, usefulness and interestingness of insights. In this paper, we propose a novel Reinforcement Learning (RL) framework for insight selection where the scoring model is trained by user feedback on interestingness and their lifestyle quality estimates. With the use of highly reusable and simple principles of automatic user simulation based on real data, we demonstrate in this preliminary study that the RL solution may improve the selection of insights towards multiple pre-defined objectives.

****

Data Leakage in Tabular Federated Learning

Mark Vero,Mislav Balunovic,Dimitar Iliev Dimitrov,Martin Vechev

While federated learning (FL) promises to preserve privacy in distributed training of deep learning models, recent work in the image and NLP domains showed that training updates leak private data of participating clients. At the same time, most high-stakes applications of FL (e.g., legal and financial) use tabular data. Compared to the NLP and image domains, reconstruction of tabular data poses several unique challenges: (i) categorical features introduce a significantly more difficult mixed discrete-continuous optimization problem, (ii) the mix of categorical and continuous features causes high variance in the final reconstructions, and (iii) structured data makes it difficult for the adversary to judge reconstruction quality. In this work, we tackle these challenges and propose the first comprehensive reconstruction attack on tabular data, called TabLeak. TabLeak is based on three key ingredients: (i) a softmax structural prior, implicitly conv erting the mixed discrete-continuous optimization problem into an easier fully continuous one, (ii) a way to reduce the variance of our reconstructions through a pooled ensembling scheme exploiting the structure of tabular data, and (iii) an entropy measure which can successfully assess reconstruction quality. Our experimental evaluation demonstrates the effectiveness of TabLeak, reaching a state-of-the-art on four popular tabular datasets. For instance, on the Adult dataset, we improve attack accuracy by 10% compared to the baseline on the practically relevant batch size of 32 and further obtain non-trivial reconstructions for batch sizes as large as 128. Our findings are important as they show that performing FL on tabular data, which often poses high privacy risks, is highly vulnerable.

****

Long-horizon video prediction using a dynamic latent hierarchy

Alexey Zakharov,Qinghai Guo,Zafeirios Fountas

The task of video prediction and generation is known to be notoriously difficult, with the research in this area largely limited to short-term predictions. Though plagued with noise and stochasticity, videos consist of features that are organised in a spatiotemporal hierarchy, different features possessing different temporal dynamics. In this paper, we introduce Dynamic Latent Hierarchy (DLH) -- a deep hierarchical latent model that represents videos as a hierarchy of latent states that evolve over separate and fluid timescales. Each latent state is a mixture distribution with two components, representing the immediate past and the predicted future, causing the model to learn transitions only between sufficiently dissimilar states, while clustering temporally persistent states closer together. Using this unique property, DLH naturally discovers the spatiotemporal structure of a dataset and learns disentangled representations across its hierarchy. We hypothesise that this simplifies the task of modeling temporal dynamics of a video, improves the learning of long-term dependencies, and reduces error accumulation. As evidence, we demonstrate that DLH outperforms state-of-the-art benchmarks in video prediction, is able to better represent stochasticity, as well as to dynamically adjust its hierarchical and temporal structure. Our paper shows, among other things, how progress in representation learning can translate into progress in prediction tasks.

****

SwinZS3: Zero-Shot Semantic Segmentation with a Swin Transformer

Tian YingJie,Wang YiQi

Zero-shot semantic segmentation (ZS3) aims at learning to classify the never-seen classes with zero training samples. Convolutional neural networks (CNNs) have recently achieved great success in this task. However, their limited attention ability constraints existing network architectures to reason based on word embeddings. In this light of the recent successes achieved by Swin Transformers, we propose SwinZS3, a new framework exploiting the visual embeddings and semantic embeddings on joint embedding space. The SwinZS3 combines a transformer image encoder with a language encoder. The image encoder is trained by pixel-text score maps using the dense language-guided semantic prototypes which are computed by the language encoder. This allows the SwinZS3 could recognize the unseen classes at test time without retraining. We experiment with our method on the ZS3 standard benchmarks (PASCAL VOC and PASCAL Context) and the results demonstrate the effectiveness of our method by showing the state-of-art performance.

****

Softened Symbol Grounding for Neuro-symbolic Systems

Zenan Li,Yuan Yao,Taolue Chen,Jingwei Xu,Chun Cao,Xiaoxing Ma,Jian L\"{u}

Neuro-symbolic learning generally consists of two separated worlds, i.e., neural network training and symbolic constraint solving,
whose success hinges on symbol grounding, a fundamental problem in AI. This paper presents a novel, softened symbol grounding process, bridging the gap between the two worlds, and resulting in an effective and efficient neuro-symbolic learning framework. Technically, the framework features (1) modeling of symbol solution states as a Boltzmann distribution, which avoids expensive state searching and facilitates mutually beneficial interactions between network training and symb

olic reasoning; (2) a new MCMC technique leveraging projection and SMT solvers, which efficiently samples from disconnected symbol solution spaces; (3) an annealing mechanism that can escape from sub-optimal symbol groundings. Experiments with three representative neuro-symbolic learning tasks demonstrate that, owing to its superior symbol grounding capability, our framework successfully solves problems well beyond the frontier of the existing proposals.

****

Encoding Recurrence into Transformers

Feiqing Huang,Kexin Lu,Yuxi CAI,Zhen Qin,Yanwen Fang,Guangjian Tian,Guodong Li

This paper novelly breaks down with ignorable loss an RNN layer into a sequence of simple RNNs, each of which can be further rewritten into a lightweight positional encoding matrix of a self-attention, named the Recurrence Encoding Matrix (REM). Thus, recurrent dynamics introduced by the RNN layer can be encapsulated into the positional encodings of a multihead self-attention, and this makes it possible to seamlessly incorporate these recurrent dynamics into a Transformer, leading to a new module, Self-Attention with Recurrence (RSA). The proposed module can leverage the recurrent inductive bias of REMs to achieve a better sample efficiency than its corresponding baseline Transformer, while the self-attention is used to model the remaining non-recurrent signals. The relative proportions of these two components are controlled by a data-driven gated mechanism, and the effectiveness of RSA modules are demonstrated by four sequential learning tasks.

****

Human-Guided Fair Classification for Natural Language Processing

Florian E. Dorner,Momchil Peychev,Nikola Konstantinov,Naman Goel,Elliott Ash,Martin Vechev

Text classifiers have promising applications in high-stake tasks such as resume screening and content moderation. These classifiers must be fair and avoid discriminatory decisions by being invariant to perturbations of sensitive attributes such as gender or ethnicity. However, there is a gap between human intuition about these perturbations and the formal similarity specifications capturing them. While existing research has started to address this gap, current methods are based on hardcoded word replacements, resulting in specifications with limited expressivity or ones that fail to fully align with human intuition (e.g., in cases of asymmetric counterfactuals). This work proposes novel methods for bridging this gap by discovering expressive and intuitive individual fairness specifications. We show how to leverage unsupervised style transfer and GPT-3's zero-shot capabilities to automatically generate expressive candidate pairs of semantically similar sentences that differ along sensitive attributes. We then validate the generated pairs via an extensive crowdsourcing study, which confirms that a lot of these pairs align with human intuition about fairness in the context of toxicity classification. Finally, we show how limited amounts of human feedback can be leveraged to learn a similarity specification that can be used to train downstream fairness-aware models.

****

Proper Scoring Rules for Survival Analysis

Hiroki Yanagisawa

Survival analysis is the problem of estimating probability distributions for future events, which can be seen as a problem in uncertainty quantification. Although there are fundamental theories on strictly proper scoring rules for uncertainty quantification, little is known about those for survival analysis. In this paper, we investigate extensions of four major strictly proper scoring rules for survival analysis. Through the extensions, we discuss and clarify the assumptions arising from the discretization of the estimation of probability distributions. We also discuss the relationship between the existing algorithms and extended scoring rules, and we propose new algorithms based on our extensions of the scoring rules for survival analysis.

****

Social Network Structure Shapes Innovation: Experience-sharing in RL with SAPIENS

Eleni Nisioti,Matéo Mahaut,Pierre-Yves Oudeyer,Ida Momennejad,Clément Moulin-Frier

The human cultural repertoire relies on innovation: our ability to continuously explore how existing elements can be combined to create new ones. Innovation is not solitary, it relies on collective accumulation and merging of previous solutions. Machine learning approaches commonly assume that fully connected multi-agent networks are best suited for innovation. However, human laboratory and field studies have shown that hierarchical innovation is more robustly achieved by dynamic social network structures. In dynamic settings, humans oscillate between innovating individually or in small clusters, and then sharing outcomes with others. To our knowledge, the role of multi-agent topology on innovation has not been systematically studied in machine learning. It remains unclear a) which social network topologies are optimal for which innovation tasks, and b) which properties of experience sharing improve multi-level innovation. Here we use a multi-level hierarchical problem setting (WordCraft), with three different innovation tasks. We systematically design networks of DQNs sharing experiences from their replay buffers in varying topologies (fully connected, small world, dynamic, ring). Comparing the level of innovation achieved by different experience-sharing topologies across different tasks shows that, first, consistent with human findings, experience sharing within a dynamic topology achieves the highest level of innovation across tasks. Second, experience sharing is not as helpful when there is a single clear path to innovation. Third, two metrics we propose, conformity and diversity of shared experience, can explain the success of different topologies on different tasks. These contributions can advance our understanding of optimal AI-AI, human-human, and human-AI collaborative networks, inspiring future tools for fostering collective innovation in large organizations.
****
Mini-batch k-means terminates within $O(d/\epsilon)$ iterations
Gregory Schwartzman
We answer the question: "Does \emph{local} progress (on batches) imply \emph{global} progress (on the entire dataset) for mini-batch k-means?". Specifically, we consider mini-batch k-means which terminates only when the improvement in the quality of the clustering on the sampled batch is below some threshold.

Although at first glance it appears that this algorithm might execute forever, we answer the above question in the affirmative and show that if the batch is of size $\tilde{\Omega}((d/\epsilon)^2)$, it must terminate within $O(d/\epsilon)$ iterations with high probability, where d is the dimension of the input, and ϵ is a threshold parameter for termination. This is true \emph{regardless} of how the centers are initialized. When the algorithm is initialized with the k-means++ initialization scheme, it achieves an approximation ratio of $O(\log k)$ (the same as the full-batch version).

Finally, we show the applicability of our results to the mini-batch k-means algorithm implemented in the scikit-learn (sklearn) python library.
****
Convergence is Not Enough: Average-Case Performance of No-Regret Learning Dynamics
Iosif Sakos,Stefanos Leonardos,William Overman,Stelios Andrew Stavroulakis,Ioannis Panageas,Georgios Piliouras
Learning in games involves two main challenges, even in settings in which agents seek to coordinate: convergence to equilibria and selection of good equilibria. Unfortunately, solving the issue of convergence, which is the focus of state-of-the-art models, conveys little information about the quality of the equilibria that are eventually reached, often none at all. In this paper, we study a class of games in which q-replicator (QRD), a widely-studied class of no-regret learning dynamics that include gradient descent, “standard” replicator, and log-barrier dynamics as special cases, can be shown to converge pointwise to Nash equilibria. This is the starting point for our main task, which is the mathematically ch

allenging problem of performance. In our main contribution, we quantify both conceptually and experimentally the outcome of optimal learning dynamics via average performance metrics, i.e., metrics that couple the regions of attraction with the quality of each attracting point. We provide an exhaustive comparison between gradient descent and “standard” replicator in a class of games with severe equilibrium selection problems and empirically extend our results to all dynamics in the QRD class. Our results combine tools from machine learning, game theory, and dynamical systems and provide a framework to initiate the systematic comparison of different optimal learning dynamics in arbitrary games.

****

Gene finding revisited: improved robustness through structured decoding from learning embeddings

Frederikke Isa Marin,Dennis Pultz,Wouter Boomsma

Gene finding is the task of identifying the locations of coding sequences within the vast amount of genetic code contained in the genome. With an ever increasing quantity of raw genome sequences, gene finding is an important avenue towards understanding the genetic information of (novel) organisms, as well as learning shared patterns across evolutionarily diverse species. The current state of the art are graphical models usually trained per organism and requiring manually curated data sets. However, these models lack the flexibility to incorporate deep learning representation learning techniques that have in recent years been transformative in the analysis of protein sequences, and which could potentially help gene finders exploit the growing number of sequenced genomes to expand performance across multiple organisms. Here, we propose a novel approach, combining learned embeddings of raw genetic sequences with exact
decoding using a latent conditional random field. We show that the model achieves performance matching the current state of the art, while increasing training robustness, and removing the need for manually fitted length distributions. As language models for DNA improve, this paves the way for more performant cross-organism gene-finders.

****

Learning Uncertainty for Unknown Domains with Zero-Target-Assumption

Yu Yu,Hassan Sajjad,Jia Xu

We introduce our Maximum-Entropy Rewarded Reinforcement Learning (MERRL) framework that selects training data for more accurate Natural Language Processing (NLP). Because conventional data selection methods select training samples based on the test domain knowledge and not on real life data, they frequently fail in unknown domains like patent and Twitter.

Our approach selects training samples that maximize information uncertainty measured by entropy, including observation entropy like empirical Shannon entropy, Min-entropy, R\'enyi entropy, and prediction entropy using mutual information, to cover more possible queries that may appear in unknown worlds. Our MERRL using regularized A2C and SAC achieves up to -99.7 perplexity decrease (-43.4\% relatively) in language modeling, +25.0 accuracy increase (+40.0\% relatively) in sentiment analysis, and +5.0 F1 score increase (+30.8\% relatively) in named entity recognition over various domains, demonstrating strong generalization power on unknown test sets.

****

Detecting Out-of-Distribution Data with Semi-supervised Graph “Feature" Networks

Debargha Ganguly,Debayan Gupta

Anomalous and out-of-distribution (OOD) data present a significant challenge to the robustness of decisions taken by deep neural networks, with myriad real-world consequences. State-of-the-art OOD detection techniques use embeddings learned by large pre-trained transformers. We demonstrate that graph structures and topological properties can be leveraged to detect both far-OOD and near-OOD data reliably, simply by characterising each data point (image) as a network of related features (visual concepts). Furthermore, we facilitate human-in-the-loop machine learning by expressing this data to comprise high-level domain-specific concepts. We obtained \textit{97.95\% AUROC} on far-OOD and \textit{98.79\% AUROC} on near-OOD detection tasks based on the LSUN dataset (comparable to the performanc e of state-of-the-art techniques).

****

Let Offline RL Flow: Training Conservative Agents in the Latent Space of Normalizing Flow

Dmitry Akimov,Vladislav Kurenkov,Alexander Nikulin,Denis Tarasov,Sergey Kolesnikov

Offline reinforcement learning aims to train a policy on a pre-recorded and fixed dataset without any additional environment interactions. There are two major challenges in this setting: (1) extrapolation error caused by approximating the value of state-action pairs not well-covered by the training data and (2) distributional shift between behavior and inference policies. One way to tackle these problems is to induce conservatism - i.e., keeping the learned policies closer to the behavioral ones. To achieve this, we build upon recent works on learning policies in latent action spaces and use a special form of normalizing flow for constructing a generative model, which we use as a conservative action encoder. This normalizing flow action encoder is pre-trained in a supervised manner on the offline dataset, and then an additional policy model - controller in the latent space - is trained via reinforcement learning. This approach avoids querying actions outside of the training dataset and therefore does not require additional regularization for out-of-dataset actions. We evaluate our method on various locomotion and navigation tasks, demonstrating that our approach outperforms recently proposed algorithms with generative action models on a large portion of datasets.

****

Towards a Complete Theory of Neural Networks with Few Neurons

Berfin Simsek,Valentin Schmutz,Wulfram Gerstner,Johanni Brea

Deep learning has seen unprecedented progress thanks to the deployment of models with millions of parameters.

On the theoretical side, an immense amount of effort has gone to understanding the dynamics of overparameterized networks.

Although now there is a well-developed theory of networks with infinitely many neurons, the classic problem of understanding how a neural network with a few neurons learns remains unsolved.

To attack this problem, we analytically study the landscapes of neural networks with few neurons.

We prove for the first time that a student network with one neuron has only one critical point --its global minimum-- when learning from a teacher network with arbitrarily many orthogonal neurons.

In addition, we prove how a neuron addition mechanism turns a minimum into a line of critical points with transitions from saddles to local minima via non-strict saddles.

Finally, we discuss how the insights we get from our novel proof techniques may shed light on the dynamics of neural networks with few neurons.

****

Machine Learning from Explanations

Jiashu Tao,Reza Shokri

Machine learning needs a huge amount of (labeled) data, as otherwise it might not learn the right model for different sub-populations, or even worse, they might pick up spurious correlations in the training data leading to brittle prediction mechanisms. Also, for small training datasets, there is a huge variability in the learned models on randomly sampled training datasets, which makes the whole process less reliable. But, collection of large amount of useful representative data, and training on large datasets, are very costly. In this paper, we present a technique to train reliable classification models on small datasets, assuming we have access to some simple explanations (e.g., subset of influential input features) on labeled data. We also propose a novel two stage training pipeline that optimizes the model's output and fine-tunes its attention in an interleaving manner, to help the model to agree with the provided explanation while learning from the data. We show that our training pipeline enables faster convergence to better models, especially when there is a severe class imbalance in the popu

lation or spurious features in the training data.
****

Functional Risk Minimization

Ferran Alet,Clement Gehring,Tomás Lozano-Pérez,Joshua B. Tenenbaum,Leslie Pack Kaelbling

In this work, we break the classic assumption of data coming from a single function $f_{\theta^*}(x)$ followed by some noise in output space $p(y|f_{\theta^*}(x))$. Instead, we model each data point (x_i,y_i) as coming from its own function f_{θ_i}. We show that this model subsumes Empirical Risk Minimization for many common loss functions, and provides an avenue for more realistic noise processes. We derive Functional Risk Minimization~(FRM), a general framework for scalable training objectives which results in better performance in small experiments in regression and reinforcement learning. We also show that FRM can be seen as finding the simplest model that memorizes the training data, providing an avenue towards understanding generalization in the over-parameterized regime.
****

Latent Linear ODEs with Neural Kalman Filtering for Irregular Time Series Forecasting

Randolf Scholz,Stefan Born,Nghia Duong-Trung,Mariano Nicolas Cruz-Bournazou,Lars Schmidt-Thieme

Over the past four years, models based on Neural Ordinary Differential Equations have become state of the art in the forecasting of irregularly sampled time series. Describing the data-generating process as a dynamical system in continuous time allows predictions at arbitrary time points. However, the numerical integration of Neural ODEs typically comes with a high computational burden or may even fail completely. We propose a novel Neural ODE model that embeds the observations into a latent space with dynamics governed by a linear ODE. Consequently, we do not require any specialized numerical integrator but only an implementation of the matrix exponential readily available in many numerical linear algebra libraries. We also introduce a novel state update component inspired by the classical Kalman filter, which, to our knowledge, makes our model the first Neural ODE variant to explicitly satisfy a specific self-consistency property. It allows forecasting irregularly sampled time series with missing values and comes with some numerical stability guarantees. We evaluate the performance on medical and climate benchmark datasets, where the model outperforms the state of the art by margins up to 30%.
****

Transformer-based model for symbolic regression via joint supervised learning

Wenqiang Li,Weijun Li,Linjun Sun,Min Wu,Lina Yu,Jingyi Liu,Yanjie Li,Songsong Tian

Symbolic regression (SR) is an important technique for discovering hidden mathematical expressions from observed data. Transformer-based approaches have been widely used for machine translation due to their high performance, and are recently highly expected to be used for SR. They input the data points, then output the expression skeleton, and finally optimize the coefficients. However, recent transformer-based methods for SR focus more attention on large scale training data and ignore the ill-posed problem: the lack of sufficient supervision, i.e., expressions that may be completely different have the same supervision because of their same skeleton, which makes it challenging to deal with data that may be from the same expression skeleton but with different coefficients. Therefore, we present a transformer-based model for SR with the ability to alleviate this problem. Specifically, we leverage a feature extractor based on pure residual MLP networks to obtain more information about data points. Furthermore, the core idea is that we propose a joint learning mechanism combining supervised contrastive learning, which makes features of data points from expressions with the same skeleton more similar so as to effectively alleviates the ill-posed problem. The benchmark results show that the proposed method is up to 25% higher with respect to the recovery rate of skeletons than typical transformer-based methods. Moreover, our method outperforms state-of-the-art SR methods based on reinforcement learning and genetic programming in terms of the coefficient of determination (R^2).

****

Gradient-Based Transfer Learning

Gustaf Tegnér,Alfredo Reichlin,Hang Yin,Hedvig Kjellstrom,Mårten Björkman,Danica Kragic

We formulate transfer learning as a meta-learning problem by extending upon the current meta-learning paradigm in that support and query data are drawn from different, but related distributions of tasks. Inspired by the success of Gradient-Based Meta-Learning (GBML), we propose to expand it to the transfer learning setting by constructing a general encoder-decoder architecture that learns a map between functionals of different domains. This is achieved by leveraging on the idea that the task-adapted parameters of a meta-learner can serve as an informative representation of the task itself. We demonstrate the proposed method on regression, prediction of dynamical systems and meta-imitation learning problems.

****

Coreset for Rational Functions

David Denisov,Ibrahim Jubran,Dan Feldman

We consider the problem of fitting a rational function $f:\mathbb{R}\to\mathbb{R}$ to a time-series $g:\{1,\cdots,n\}\to\mathbb{R}$. This is by minimizing the sum of distances (loss function) $\ell(f):=\sum_{i=1}^n |f(i)-g(i)|$, possibly with additional constraints and regularization terms that may depend on f. Our main motivation is to approximate such a time-series by a recursive sequence model $F_n=\sum_{i=1}^k \theta_i F_{n-i}$, e.g. a Fibonacci sequence, where $\theta\in \mathbb{R}^k$ are the model parameters, and $k\geq1$ is constant.

For $\varepsilon\in(0,1)$, an ε-coreset for this problem is a small data structure that approximates $\ell(g)$ up to $1\pm\varepsilon$ multiplicative factor, for every rational function g of constant degree.

We prove that every signal has an ε-coreset of size $O(n^{0.001}/\varepsilon^2)$, and provide a construction algorithm that computes it in $O(n^{1.001})$ time.

Open source code is provided, as well as extensive experimental results, on both real and synthetic datasets, which compare our method to existing solvers from Scipy.

****

Transformer needs NMDA receptor nonlinearity for long-term memory

Dong-Kyum Kim,Jea Kwon,Meeyoung Cha,C. Justin Lee

The NMDA receptor (NMDAR) in the hippocampus is essential for learning and memory. We find an interesting resemblance between deep models' nonlinear activation function and the NMDAR's nonlinear dynamics. In light of a recent study that compared the transformer architecture to the formation of hippocampal memory, this paper presents new findings that NMDAR-like nonlinearity may be essential for consolidating short-term working memory into long-term reference memory. We design a navigation task assessing these two memory functions and show that manipulating the activation function (i.e., mimicking the Mg^{2+}-gating of NMDAR) disrupts long-term memory formation. Our experimental data suggest that the concept of place cells and reference memory may reside in the feed-forward network layer of transformers and that nonlinearity plays a key role in these processes. Our findings propose that the transformer architecture and hippocampal spatial representation resemble by sharing the overlapping concept of NMDAR-like nonlinearity.

****

Simple Spectral Graph Convolution from an Optimization Perspective

Hao Zhu,Piotr Koniusz

Recent studies on SGC, PageRank and S^{2}GC have demonstrated that several graph diffusion techniques are straightforward, quick, and effective for tasks in the graph domain like node classification. Even though these techniques do not even need labels, they can nevertheless produce more discriminating features than raw attributes for downstream tasks with different classifiers. These methods are data-independent and thus primarily rely on some empirical parameters on polynomial bases (e.g., Monomial and Chebyshev), which ignore the homophily of graphs and the attribute distribution. They are more insensitive to heterophilous graphs due to the low-pass filtering. Although there are many approache s focusing on GNNs based on heterophilous graphs, these approaches are dependent on label information to learn model parameters. In this paper, we study the question: are labels a necessity for GNNs with heterophilous graphs? Based on this question, we propose a framework of self-representation on graphs related to the Least Squares problem. Specifically, we use Generalized Minimum RESidual (GMRES) method, which finds the least squares solution over Krylov subspaces. In theoretical analysis, without label information, we enjoy better features with graph convolution.

The proposed method, like previous data-independent methods, is not a deep model and is, therefore, quick, scalable, and simple. We also show performance guarantees for models on real and synthetic data. On a benchmark of real-world datasets, empirically, our method is competitive with existing deep models for node classification.

****

QAID: Question Answering Inspired Few-shot Intent Detection

Asaf Yehudai,Matan Vetzler,Yosi Mass,Koren Lazar,Doron Cohen,Boaz Carmeli

Intent detection with semantically similar fine-grained intents is a challenging task. To address it, we reformulate intent detection as a question-answering retrieval task by treating utterances and intent names as questions and answers. To that end, we utilize a question-answering retrieval architecture and adopt a two stages training schema with batch contrastive loss. In the pre-training stage, we improve query representations through self-supervised training. Then, in the fine-tuning stage, we increase contextualized token-level similarity scores between queries and answers from the same intent. Our results on three few-shot intent detection benchmarks achieve state-of-the-art performance.

****

Rethinking the Value of Prompt Learning for Vision-Language Models

Peisong Wang,Weihan Chen,Weixiang Xu,Qinghao Hu,Jian Cheng

Large-scale visual-language pre-training like CLIP has demonstrated great success in open-set visual concept learning that enables zero-shot transfer to downstream tasks through prompting. To automate prompt engineering, prompt learning is proposed to automatically learn the optimal task-relevant prompts. In this paper, we make some surprising observations that contradict common beliefs about prompts. We observe that even random prompts can achieve pretty good performance for zero-shot recognition. We also find that prompt learning gives comparable or worse performance than directly fine-tuning of the linear classifier. Moreover, prompt learning is no more than parameter-efficient learning, and is a trade-off between optimality and generalization. Our results highlight the need for the rethinking of existing prompt learning, more careful baseline evaluations in future research on prompt learning methods in vision-language models.

****

Disentangled Feature Swapping Augmentation for Weakly Supervised Semantic Segmentation

Junehyoung Kwon,Eunju Lee,YunSung Cho,YoungBin Kim

Weakly supervised semantic segmentation utilizes a localization map obtained from a classifier to generate a pseudo-mask. However, classifiers utilize background cues to predict class labels because of a biased dataset consisting of images, in which specific objects frequently co-occur with certain backgrounds. Consequently, the classifier confuses the background with the target objects, resulting in inaccurate localization maps. To this end, we propose DisEntangled FeaTure swapping augmentation(DEFT) to prevent the classifier from being biased by a misleading correlation. Our method first disentangles the foreground and background features. Then, we randomly swap the disentangled features within mini-batches via a two-way process. These features contain various contexts that do not appear in the biased dataset, but the class relevant representation is preserved. In addition, we introduce training schemes to obtain further performance gains. Experimental results showed that when our augmentation was used in various weakly supervised semantic segmentation methods trained on the Pascal VOC 2012 dataset, the performance of the localization maps and pseudo-mask as well as the segmentation results improved.

****
Distributed Least Square Ranking with Random Features
Rong Yin,Yong Liu,Weiping Wang,Dan Meng
In this paper, we study the statistical properties of pairwise ranking using distributed learning and random features (called DRank-RF) and establish its convergence analysis in probability. Theoretical analysis shows that DRank-RF remarkably reduces the computational requirements while preserving a satisfactory convergence rate. An extensive experiment verifies the effectiveness of DRank-RF. Furthermore, to improve the learning performance of DRank-RF, we propose an effective communication strategy for it and demonstrate the power of communications via theoretical assessments and numerical experiments.
****
Doing Fast Adaptation Fast: Conditionally Independent Deep Ensembles for Distribution Shifts
Wanqian Yang,Aahlad Manas Puli,Andrew Gordon Wilson,Rajesh Ranganath
Classifiers in a diverse ensemble capture distinct predictive signals, which is valuable for datasets containing multiple strongly predictive signals. Performing fast adaptation at test time allows us to generalize to distributions where certain signals are no longer predictive, or to avoid relying on sensitive or protected attributes. However, ensemble learning is often expensive, even more so when we need to enforce diversity constraints between the high-dimensional representations of the classifiers. Instead, we propose an efficient and fast method for learning ensemble diversity. We minimize conditional mutual information of the output distributions between classifiers, a quantity which can be cheaply and exactly computed from empirical data. The resulting ensemble contains individually strong predictors that are only dependent because they predict the label. We demonstrate the efficacy of our method on shortcut learning tasks. Performing fast adaptation on our ensemble selects shortcut-invariant models that generalize well to test distributions where the shortcuts are uncorrelated with the label.

****
Solving stochastic weak Minty variational inequalities without increasing batch size
Thomas Pethick,Olivier Fercoq,Puya Latafat,Panagiotis Patrinos,Volkan Cevher
This paper introduces a family of stochastic extragradient-type algorithms for a class of nonconvex-nonconcave problems characterized by the weak Minty variational inequality (MVI). Unlike existing results on extragradient methods in the monotone setting, employing diminishing stepsizes is no longer possible in the weak MVI setting. This has led to approaches such as increasing batch sizes per iteration which can however be prohibitively expensive. In contrast, our proposed methods involves two stepsizes and only requires one additional oracle evaluation per iteration. We show that it is possible to keep one fixed stepsize while it is only the second stepsize that is taken to be diminishing, making it interesting even in the monotone setting. Almost sure convergence is established and we provide a unified analysis for this family of schemes which contains a nonlinear generalization of the celebrated primal dual hybrid gradient algorithm.

****
Diversity Boosted Learning for Domain Generalization with a Large Number of Domains
XI LENG,Xiaoying Tang,Yatao Bian
Machine learning algorithms minimizing the average training loss typically suffer from poor generalization performance. It inspires various works for domain generalization (DG), among which a series of methods work by $O(n^2)$ pairwise domain operations with n domains, where each one is often costly. Moreover, while a common objective in the DG literature is to learn invariant representations against spurious correlations induced by domains, we point out the insufficiency of it and highlight the importance of alleviating spurious correlations caused by objects. Based on the observation that diversity helps mitigate spurious correlations, we propose a Diversity boosted twO-level saMplIng framework (DOMI) to effi

ciently sample the most informative ones among a large number of domains and data points. We show that DOMI helps train robust models against spurious correlations from both domain-side and object-side, substantially enhancing the performance of five backbone DG algorithms on Rotated MNIST and Rotated Fashion MNIST.

****

Towards Performance-maximizing Network Pruning via Global Channel Attention

Yingchun Wang,Linchuan Xu,Song Guo,Jingcai Guo,Weizhan Zhang,Jie Zhang,shuai chen

Network pruning has attracted increasing attention recently for its capability of transferring large-scale neural networks (e.g., CNNs) into resource-constrained devices. Such a transfer is typically achieved by removing redundant network parameters while retaining its generalization performance in a static or dynamic pruning manner. Concretely, static pruning usually maintains a larger and fit-to-all (samples) compressed network by removing the same channels for all samples, while dynamic pruning can adaptively remove (more) different channels for different samples and obtain state-of-the-art performance along with a higher compression ratio. However, since the system has to preserve the complete network information for sample-specific pruning, dynamic pruning methods are usually not memory-efficient. In this paper, our interest is to explore a static alternative, dubbed GlobalPru, to conventional static pruning methods that can take into account both compression ratio and model performance maximization. Specifically, a novel channel attention-based learn-to-rank algorithm is proposed to learn the global channel attention of the network for various samples, wherein, each sample-specific channel saliency is forced to reach an agreement on the global ranking. Hence, all samples can empirically share the same pruning priority of channels to achieve channel pruning with minimal performance loss. Extensive experiments demonstrate that the proposed GlobalPru can achieve better performance than state-of-the-art static and dynamic pruning methods by significant margins.

****

Adaptive Block-wise Learning for Knowledge Distillation

Tianyi Lei,Junyu Xie,Wang qian,Dezhong Peng,Xu Wang

Knowledge distillation allows the student network to improve its performance under the supervision of transferred knowledge. Existing knowledge distillation methods are implemented under the implicit hypothesis that knowledge from teacher and student contributes to each layer of the student network to the same extent. In this work, we argue that there should be different contributions of knowledge from the teacher and the student during training for each layer. Experimental results evidence this argument. To the end, we propose a novel Adaptive Block-wise Learning~(ABL) for Knowledge Distillation to automatically balance teacher-guided knowledge between self-knowledge in each block. Specifically, to solve the problem that the error backpropagation algorithm cannot assign weights to each block of the student network independently, we leverage the local error signals to approximate the global error signals on student objectives. Moreover, we utilize a set of meta variables to control the contribution of the student knowledge and teacher knowledge to each block during the training process. Finally, the extensive experiments prove the effectiveness of our method. Meanwhile, ABL provides an insightful view that in the shallow blocks, the weight of teacher guidance is greater, while in the deep blocks, student knowledge has more influence.

****

Curriculum-based Co-design of Morphology and Control of Voxel-based Soft Robots

Yuxing Wang,Shuang Wu,Haobo Fu,QIANG FU,Tiantian Zhang,Yongzhe Chang,Xueqian Wang

Co-design of morphology and control of a Voxel-based Soft Robot (VSR) is challenging due to the notorious bi-level optimization. In this paper, we present a Curriculum-based Co-design (CuCo) method for learning to design and control VSRs through an easy-to-difficult process. Specifically, we expand the design space from a small size to the target size gradually through a predefined curriculum. At each learning stage of the curriculum, we use reinforcement learning to simultaneously train the design policy and the control policy, which is enabled by incorporating the design process into the environment and using differentiable policy representations. The converged morphology and the learned policies from last stage are inherited and then serve as the starting point for the next stage. In empirical studies, we show that CuCo is more efficient in creating larger robots with better performance by reusing the practical design and control patterns learned within each stage, in comparison to prior approaches that learn from scratch in the space of target size.

****

Object-Centric Learning with Slot Mixture Models

Daniil Kirilenko,Alexey Kovalev,Aleksandr Panov

Object-centric architectures usually apply some differentiable module on the whole feature map to decompose it into sets of entities representations called slots. Some of these methods structurally resemble clustering algorithms, where the center of the cluster in latent space serves as slot representation. Slot Attention is an example of such a method as a learnable analog of the soft k-Means algorithm. In our work, we use the learnable clustering method based on Gaussian Mixture Model, unlike other approaches we represent slots not only as centers of clusters but we also use information about the distance between clusters and assigned vectors, which leads to more expressive slots representations. Our experiments demonstrate that using this approach instead of Slot Attention improves performance in different scenarios achieving state-of-the-art performance in the set property prediction task.

****

WiNeRT: Towards Neural Ray Tracing for Wireless Channel Modelling and Differentiable Simulations

Tribhuvanesh Orekondy,Pratik Kumar,Shreya Kadambi,Hao Ye,Joseph Soriaga,Arash Behboodi

In this paper, we work towards a neural surrogate to model wireless electro-magnetic propagation effects in indoor environments.

Such neural surrogates provide a fast, differentiable, and continuous representation of the environment and enables end-to-end optimization for downstream tasks (e.g., network planning). Specifically, the goal of the paper is to render the wireless signal (e.g., time-of-flights, power of each path) in an environment as a function of the sensor's spatial configuration (e.g., placement of transmit and receive antennas). NeRF-based approaches have shown promising results in the visual setting (RGB image signal, with a camera sensor), where the key idea is to algorithmically evaluate the 'global' signal (e.g., using volumetric rendering) by breaking it down in a sequence of 'local' evaluations (e.g., using co-ordinate neural networks). In a similar spirit, we model the time-angle channel impulse response (the global wireless signal) as a superposition of multiple paths. The wireless characteristics (e.g., power) of each path is a result of multiple evaluations of a neural network that learns implicit ray-surface interaction properties. We evaluate our approach in multiple indoor scenarios and demonstrate that our model achieves strong performance (e.g., $<$0.33ns error in time-of-flight predictions). Furthermore, we demonstrate that our neural surrogate whitens the `black-box' wireless simulators, and thus enables inverse rendering applications (e.g., user localization).

****

Pocket-specific 3D Molecule Generation by Fragment-based Autoregressive Diffusion Models

Xingang Peng,Jiaqi Guan,Jian Peng,Jianzhu Ma

Autoregressive model is widely adopted to generate 3D molecules which can fit any protein binding pocket. Current autoregressive model suffers from two major drawbacks. First, it is hard to capture local geometric patterns as only one atom is generated at each step. Second, most of the autoregressive models generate atoms and chemical bonds in two separate processes, which causes a number of problems such as incorrect counts of rings, a bias distribution of bond lengths, and inaccurate 3D molecular structures. To tackle this problem, we designed a model, named FragDiff, to generate 3D molecules fragment-by-fragment for pockets. In each generation step, FragDiff places a molecular fragment around the pocket by using E(3)-equivariant diffusion generative models to simultaneously predict the atom types, atom coordinates and the chemical bonds of the fragment. Extensive experimental results confirm our assumption that unifying the atoms and bonds generations could significantly improve the quality of the sampled 3D molecules in terms of more accurate distributions of 2D subgraphs and 3D substructures.

****

Towards scalable and non-IID robust Hierarchical Federated Learning via Label-driven Knowledge Aggregator

Duong Minh Nguyen,Viet Quoc Pham,Hoang Thai Dinh,Diep Nguyen,Long Tran-Thanh,Won-Joo Hwang

In real-world applications, Federated Learning (FL) meets two challenges: (1) scalability, especially when applied to massive IoT networks, and (2) how to be robust against an environment with heterogeneous data. Realizing the first problem, we aim to design a novel FL framework named Full-stack FL (F2L). More specifically, F2L utilizes a hierarchical network architecture, making extending the FL network accessible without reconstructing the whole network system. Moreover, leveraging the advantages of hierarchical network design, we propose a new label-driven knowledge distillation (LKD) technique at the global server to address the second problem. As opposed to current knowledge distillation techniques, LKD is capable of training a student model, which consists of good knowledge from all teachers' models. Therefore, our proposed algorithm can effectively extract the knowledge of the regions' data distribution (i.e., the regional aggregated models) to reduce the divergence between clients' models when operating under the FL system with non-independent identically distributed data. Extensive experiment results reveal that: (i) our F2L method can significantly improve the overall FL efficiency in all global distillations, and (ii) F2L rapidly achieves convergence as global distillation stages occur instead of increasing on each communication cycle.

****

LS-IQ: Implicit Reward Regularization for Inverse Reinforcement Learning

Firas Al-Hafez,Davide Tateo,Oleg Arenz,Guoping Zhao,Jan Peters

Recent methods for imitation learning directly learn a Q-function using an implicit reward formulation rather than an explicit reward function. However, these methods generally require implicit reward regularization to improve stability and often mistreat absorbing states. Previous works show that a squared norm regularization on the implicit reward function is effective, but do not provide a theoretical analysis of the resulting properties of the algorithms. In this work, we show that using this regularizer under a mixture distribution of the policy and the expert provides a particularly illuminating perspective: the original objective can be understood as squared Bellman error minimization, and the corresponding optimization problem minimizes a bounded χ^2-Divergence between the expert and the mixture distribution. This perspective allows us to address instabilities and properly treat absorbing states. We show that our method, Least Squares Inverse Q-Learning (LS-IQ), outperforms state-of-the-art algorithms, particularly in environments with absorbing states. Finally, we propose to use an inverse dynamics model to learn from observations only. Using this approach, we retain performance in settings where no expert actions are available.

****

Humanly Certifying Superhuman Classifiers

Qiongkai Xu,Christian Walder,Chenchen Xu

This paper addresses a key question in current machine learning research: if we believe that a model's predictions might be better than those given by human experts, how can we (humans) verify these beliefs? In some cases, this ``superhuman'' performance is readily demonstrated; for example by defeating top-tier human players in traditional two player games. On the other hand, it can be challenging to evaluate classification models that potentially surpass human performance. Indeed, human annotations are often treated as a ground truth, which implicitly assumes the superiority of the human over any models trained on human annotations. In reality, human annotators are subjective and can make mistakes. Evaluating the performance with respect to a genuine oracle is more objective and reliable, even when querying the oracle is more expensive or sometimes impossible. In th is paper, we first raise the challenge of evaluating the performance of both humans and models with respect to an oracle which is $\textit{unobserved}$. We develop a theory for estimating the accuracy compared to the oracle, using only imperfect human annotations for reference. Our analysis provides an executable recipe for detecting and certifying superhuman performance in this setting, which we believe will assist in understanding the stage of current research on classification. We validate the convergence of the bounds and the assumptions of our theory on carefully designed toy experiments with known oracles. Moreover, we demonstrate the utility of our theory by meta-analyzing large-scale natural language processing tasks, for which an oracle does not exist, and show that under our mild assumptions a number of models from recent years have already achieved superhuman performance with high probability---suggesting that our new oracle based performance evaluation metrics are overdue as an alternative to the widely used accuracy metrics that are naively based on imperfect human annotations.

****

Share Your Representation Only: Guaranteed Improvement of the Privacy-Utility Tradeoff in Federated Learning

Zebang Shen,Jiayuan Ye,Anmin Kang,Hamed Hassani,Reza Shokri

Repeated parameter sharing in federated learning causes significant information leakage about private data, thus defeating its main purpose: data privacy. Mitigating the risk of this information leakage, using state of the art differentially private algorithms, also does not come for free. Randomized mechanisms can prevent convergence of models on learning even the useful representation functions, especially if there is more disagreement between local models on the classification functions (due to data heterogeneity). In this paper, we consider a representation federated learning objective that encourages various parties to collaboratively refine the consensus part of the model, with differential privacy guarantees, while separately allowing sufficient freedom for local personalization (without releasing it). We prove that in the linear representation setting, while the objective is non-convex, our proposed new algorithm \DPFEDREP\ converges to a ball centered around the \emph{global optimal} solution at a linear rate, and the radius of the ball is proportional to the reciprocal of the privacy budget. With this novel utility analysis, we improve the SOTA utility-privacy trade-off for this problem by a factor of $\sqrt{d}$, where d is the input dimension. We empirically evaluate our method with the image classification task on CIFAR10, CIFAR100, and EMNIST, and observe a significant performance improvement over the prior work under the same small privacy budget. The code can be found in this link, https://github.com/shenzebang/CENTAUR-Privacy-Federated-Representation-Learning.

****

Quantized Disentangled Representations for Object-Centric Visual Tasks

Daniil Kirilenko,Alexandr Korchemnyi,Alexey Kovalev,Aleksandr Panov

Recently, the pre-quantization of image features into discrete latent variables has helped to achieve remarkable results in image modeling. In this paper, we propose a method to learn discrete latent variables applied to object-centric tasks. In our approach, each object is assigned a slot which is represented as a vector generated by sampling from non-overlapping sets of low-dimensional discrete variables.

We empirically demonstrate that embeddings from the learned discrete latent spaces have the disentanglement property. The model is trained with a set prediction and object discovery as downstream tasks. It achieves the state-of-the-art results on the CLEVR dataset among a class of object-centric methods for set prediction task. We also demonstrate manipulation of individual objects in a scene with controllable image generation in the object discovery setting.

****

Supervised Random Feature Regression via Projection Pursuit

Jingran Zhou,Ling Zhou,shaogao lv

Random feature methods and neural network models are two popular nonparametric modeling methods, which are regarded as representatives of shallow learning and Neural Network, respectively. In practice random feature methods are short of th

e capacity of feature learning, while neural network methods lead to computationally heavy problems. This paper aims at proposing a flexible but computational efficient method for general nonparametric problems. Precisely, our proposed method is a feed-forward two-layer nonparametric estimation, and the first layer is used to learn a series of univariate basis functions for each projection variable, and then search for their optimal linear combination for each group of these learnt functions. Based on all the features derived in the first layer, the second layer attempts at learning a single index function with an unknown activation function. Our nonparametric estimation takes advantage of both random features and neural networks, and can be seen as an intermediate bridge between them.
****
Graph Spline Networks for Efficient Continuous Simulation of Dynamical Systems
Chuanbo Hua,Federico Berto,Michael Poli,Stefano Massaroli,Jinkyoo Park
While complex simulations of physical systems have been widely studied in engineering and scientific computing, lowering their often prohibitive computational requirements has only recently been tackled by deep learning approaches. In this paper, we present GraphSplineNets, a novel deep learning approach to speed up simulation of physical systems with spatio-temporal continuous outputs by exploiting the synergy between graph neural networks (GNN) and orthogonal spline collocation (OSC). Two differentiable time-oriented OSC and spatial-oriented OSC are applied to bridge the gap between discrete GNN outputs and generate continuous solutions at any location in space and time without explicit prior knowledge of underlying differential equations. Moreover, we introduce an adaptive collocation strategy in space to enable the model to sample from the most important regions. Our model improves on widely used graph neural networks for physics simulation on both efficiency and solution accuracy. We demonstrate SplineGraphNets in predicting complex dynamical systems such as the heat equation, damped wave propagation and the Navier-Stokes equations for incompressible flow, where they improve accuracy of more than 25% while providing at least 60% speedup.
****
Online black-box adaptation to label-shift in the presence of conditional-shift
Faruk Ahmed,Aaron Courville
We consider an out-of-distribution setting where trained predictive models are deployed online in new locations (inducing conditional-shift), such that these locations are also associated with differently skewed target distributions (label-shift). While approaches for online adaptation to label-shift have recently been discussed by Wu et al. (2021), the potential presence of concurrent conditional-shift has not been considered in the literature, although one might anticipate such distributional shifts in realistic deployments. In this paper, we empirically explore the effectiveness of online adaptation methods in such situations on three synthetic and two realistic datasets, comprising both classification and regression problems. We show that it is possible to improve performance in these settings by learning additional hyper-parameters to account for the presence of conditional-shift by using appropriate validation sets.
****
RuDar: Weather Radar Dataset for Precipitation Nowcasting with Geographical and Seasonal Variability
Petr Vytovtov,Eugenia Elistratova,Evgenii Tsymbalov,Alexander Ganshin,Yuri Pavlyukov
Precipitation nowcasting, a short-term (up to six hours) rain prediction, is arguably one of the most demanding weather forecasting tasks.
To achieve accurate predictions, a forecasting model should consider miscellaneous meteorological and geographical data sources.
Currently available datasets provide information only about precipitation intensity, vertically integrated liquid (VIL), or maximum reflectivity on the vertical section.
Such single-level or aggregated data lacks description of the reflectivity change in vertical dimension, simplifying or distorting the corresponding models.

To fill this gap, we introduce an additional dimension of the precipitation meas urements in the RuDar dataset that incorporates 3D radar echo observations.

Measurements are collected from 30 weather radars located mostly in the European part of Russia, covering multiple climate zones.

Radar product updates every 10 minutes with a 2 km spatial resolution.

The measurements include precipitation intensity (mm/h) at an altitude of 600 m, reflectivity (dBZ) and radial velocity (m/s) at 10 altitude levels from 1 km to 10 km with 1 km step.

We also add the orography information as it affects the intensity and distribution of precipitation.

The dataset includes over 50 000 timestamps over a two-year period from 2019 to 2021, totalling in roughly 100 GB of data.

We evaluate several baselines, including optical flow and neural network models, for precipitation nowcasting on the proposed data. We also evaluate the uncertainty quantification for the ensemble scenario and show that the corresponding estimates do correlate with the ensemble errors on different sections of data.

We believe that RuDar dataset will become a reliable benchmark for precipitation nowcasting models and also will be used in other machine learning tasks, e.g., in data shift studying, anomaly detection, or uncertainty estimation.

Both dataset and code for data processing and model preparation are publicly available.

****

Learning Representations for Reinforcement Learning with Hierarchical Forward Models

Trevor McInroe,Lukas Schäfer,Stefano V Albrecht

Learning control from pixels is difficult for reinforcement learning (RL) agents because representation learning and policy learning are intertwined. Previous approaches remedy this issue with auxiliary representation learning tasks, but they either do not consider the temporal aspect of the problem or only consider single-step transitions, which may miss relevant information if important environmental changes take many steps to manifest. We propose Hierarchical k-Step Latent (HKSL), an auxiliary task that learns representations via a hierarchy of forward models that operate at varying magnitudes of step skipping while also learning to communicate between levels in the hierarchy. We evaluate HKSL in a suite of 30 robotic control tasks with and without distractors and a task of our creation. We find that HKSL either converges to higher episodic returns or optimal performance more quickly than several current baselines. Furthermore, we find that HKSL's representations capture task-relevant details accurately across timescales (even in the presence of distractors) and that communication channels between hierarchy levels organize information based on both sides of the communication process, both of which improve sample efficiency.

****

xTrimoABFold: Improving Antibody Structure Prediction without Multiple Sequence Alignments

Yining Wang,Xumeng Gong,Shaochuan Li,Bing Yang,YiWu Sun,Chuan Shi,Hui Li,Yangang Wang,Cheng Yang,Le Song

Antibody, used by the immune system to identify and neutralize foreign objects such as pathogenic bacteria and viruses, plays an important role in immune system . In the field of drug engineering, the essential task is designing a novel antibody to make sure its paratope (substructures in the antibody) binds to the epitope of the specific antigen with high precision. Also, understanding the structure of antibody and its paratope can facilitate a mechanistic understanding of the function. Therefore, antibody structure prediction has always been a highly valuable problem for drug discovery. AlphaFold2, a breakthrough in the field of structural biology, provides a feasible solution to predict protein structure based on protein sequences and computationally expensive coevolutionary multiple sequence alignments (MSAs). However, the computational efficiency and undesirable prediction accuracy on antibody, especially on the complementarity-determining regions (CDRs) of antibody limit its applications on the industrially high-throughput drug design. In this paper, we present a novel method named xTrimoABFold to

predict antibody structure from antibody sequence based on a pretrained antibody language model (ALM) as well as homologous templates, which are searched from protein database (PDB) via fast and cheap algorithms. xTrimoABFold outperforms the MSA-based AlphaFold2 and the protein language model based SOTAs, e.g., OmegaFold, HelixFold-Single and IgFold with a large significant margin (30+% improvement on RMSD) while performs 151x faster than AlphaFold2. To the best of our knowledge, xTrimoABFold is the best antibody structure predictor to date in the world.
****

Thresholded Lexicographic Ordered Multi-Objective Reinforcement Learning

Alperen Tercan,Vinayak Prabhu

Lexicographic multi-objective problems, which impose a lexicographic importance order over the objectives, arise in many real-life scenarios. Existing Reinforcement Learning work directly addressing lexicographic tasks has been scarce. The few proposed approaches were all noted to be heuristics without theoretical guarantees as the Bellman equation is not applicable to them. Additionally, the practical applicability of these prior approaches also suffers from various issues such as not being able to reach the goal state. While some of these issues have been known before, in this work we investigate further shortcomings, and propose fixes for improving practical performance in many cases. We also present a policy optimization approach using our Lexicographic Projection Optimization (LPO) algorithm that has the potential to address these theoretical and practical concerns. Finally, we demonstrate our proposed algorithms on benchmark problems.
****

HOW SAMPLING AFFECTS TRAINING: AN EFFECTIVE SAMPLING THEORY STUDY FOR LONG-TAILED IMAGE CLASSIFICATION

Gong Zhang,Yongqiang Gao,Liu Haijing

The long-tailed image classification problem has been very challenging for a longtime. Suffered from the unbalanced distribution of categories, many deep vision classification methods perform well in the head classes while poor in the tail ones. This paper proposes an effective sampling theory, attempting to provide a theoretical explanation for the decoupling representation and classifier for long-tailed image classification. To apply the above sampling theory in practice, a general jitter sampling strategy is proposed. Experiments show that variety of long-tailed distribution algorithms exhibit better performance based on the effective sampling theory. The code will be released soon later.
****

EquiMod: An Equivariance Module to Improve Visual Instance Discrimination

Alexandre DEVILLERS,Mathieu Lefort

Recent self-supervised visual representation methods are closing the gap with supervised learning performance. Most of these successful methods rely on maximizing the similarity between embeddings of related synthetic inputs created through data augmentations. This can be seen as a task that encourages embeddings to leave out factors modified by these augmentations, i.e. to be invariant to them. However, this only considers one side of the trade-off in the choice of the augmentations: they need to strongly modify the images to avoid simple solution shortcut learning (e.g. using only color histograms), but on the other hand, augmentations-related information may be lacking in the representations for some downstream tasks (e.g. literature shows that color is important for bird and flower classification). Few recent works proposed to mitigate this problem of using only an invariance task by exploring some form of equivariance to augmentations. This has been performed by learning additional embeddings space(s), where some augmentation(s) cause embeddings to differ, yet in a non-controlled way. In this work, we introduce EquiMod a generic equivariance module that structures the learned latent space, in the sense that our module learns to predict the displacement in the embedding space caused by the augmentations. We show that applying that module to state-of-the-art invariance models, such as BYOL and SimCLR, increases the performances on the usual CIFAR10 and ImageNet datasets. Moreover, while our model could collapse to a trivial equivariance, i.e. invariance, we observe that it instead automatically learns to keep some augmentations-related information beneficial to the representations.

****

Manipulating Multi-agent Navigation Task via Emergent Communications

Han Yu,Hengtong Lu,Caixia Yuan,Xiaojie Wang

Multi-agent corporations struggle to efficiently sustain grounded communications with a specific task goal. Existing approaches are limited in their simple task settings and single-turn communications. This work describes a multi-agent communication scenario via emergent language in a navigation task. This task involves two agents with unequal abilities: the tourist (agent A) who can only observe the surroundings and the guide (agent B) who has a holistic view but does not know the initial position of agent A. They communicate with the emerged language grounded through the environment and a common task goal: to help the tourist find the target place. We release a new dataset of 3000 scenarios that involve multi-agent visual and language navigation. We also seek to address the multi-agent emergent communications by proposing a collaborative learning framework that enables the agents to generate and understand emergent language and solve tasks. The framework is trained with reinforcement learning by maximizing the task success rate in an end-to-end manner. Results show that the proposed framework achieves competing performance in both the accuracy of language understanding and the task success rate. We also discuss the explanations of the emerged language.

****

Task-Aware Information Routing from Common Representation Space in Lifelong Learning

Prashant Shivaram Bhat,Bahram Zonooz,Elahe Arani

Intelligent systems deployed in the real world suffer from catastrophic forgetting when exposed to a sequence of tasks. Humans, on the other hand, acquire, consolidate, and transfer knowledge between tasks that rarely interfere with the consolidated knowledge. Accompanied by self-regulated neurogenesis, continual learning in the brain is governed by the rich set of neurophysiological processes that harbor different types of knowledge which are then integrated by the conscious processing. Thus, inspired by Global Workspace Theory of conscious information access in the brain, we propose TAMiL, a continual learning method that entails task-attention modules to capture task-specific information from the common representation space. We employ simple, undercomplete autoencoders to create a communication bottleneck between the common representation space and the global workspace, allowing only the task-relevant information to the global workspace, thereby greatly reducing task interference. Experimental results show that our method outperforms state-of-the-art rehearsal-based and dynamic sparse approaches and bridges the gap between fixed capacity and parameter isolation approaches while being scalable. We also show that our method effectively mitigates catastrophic forgetting while being well-calibrated with reduced task-recency bias.

****

CodeBPE: Investigating Subtokenization Options for Large Language Model Pretraining on Source Code

Nadezhda Chirkova,Sergey Troshin

Recent works have widely adopted large language model pretraining for source code, suggested source code-specific pretraining objectives and investigated the applicability of various Transformer-based language model architectures for source code. This work investigates another important aspect of such models, the effect of different subtokenization options, and aims at identifying most effective and length-efficient subtokenizations, taking into account source code specifics. We propose subtokenziation that reduces average length by 17--40% without downstream performance drop, and show that a carefully chosen subtokenization may improve quality by 0.5-2%, possibly with some length increase.

****

Transport with Support: Data-Conditional Diffusion Bridges

Ella Tamir,Martin Trapp,Arno Solin

The dynamic Schrödinger bridge problem provides an appealing setting for posing optimal transport problems as learning non-linear diffusion processes and enables efficient iterative solvers. Recent works have demonstrated state-of-the-art results (eg, in modelling single-cell embryo RNA sequences or sampling from compl ex posteriors) but are typically limited to learning bridges with only initial and terminal constraints. Our work extends this paradigm by proposing the Iterative Smoothing Bridge (ISB). We combine learning diffusion models with Bayesian filtering and optimal control, allowing for constrained stochastic processes governed by sparse observations at intermediate stages and terminal constraints. We assess the effectiveness of our method on synthetic and real-world data and show that the ISB generalises well to high-dimensional data, is computationally efficient, and provides accurate estimates of the marginals at intermediate and terminal times.

****

Randomized Sharpness-Aware Training for Boosting Computational Efficiency in Deep Learning

Yang Zhao

By driving optimizers to converge to flat minima, sharpness-aware learning algorithms (such as SAM) have shown the power to achieve state-of-art performances. However, these algorithms will generally incur one extra forward-backward propagation at each training iteration, which largely burdens the computation especially for scalable models. To this end, we propose an efficient training scheme, called Randomized Sharpness-Aware Training (RST). Optimizers in RST would perform a Bernoulli trial at each iteration to choose randomly from base algorithms (SGD) and sharpness-aware algorithms (SAM) with a probability arranged by a predefined scheduling function. Due to the mixture of base algorithms, the overall count of propagation pairs could be largely reduced. Also, we give theoretical analysis on the convergence of RST. Then, we empirically study the computation cost and effect of various types of scheduling functions, and give directions on setting appropriate scheduling functions. Further, we extend the RST to a general framework (G-RST), where we can adjust regularization degree on sharpness freely for any scheduling function. We show that G-RST can outperform SAM in most cases while saving 50\% extra computation cost.

****

Self-Supervised Off-Policy Ranking via Crowd Layer

Pengjie Gu,Mengchen Zhao,Jianye HAO,Bo An

Off-policy evaluation (OPE) aims to estimate the online performance of target policies given dataset collected by some behavioral policies. OPE is crucial in many applications where online policy evaluation is expensive. However, existing OPE methods are far from reliable. Fortunately, in many real-world scenarios, we care only about the ranking of the evaluating policies, rather than their exact online performance. Existing works on off-policy ranking (OPR) adopt a supervised training paradigm, which assumes that there are plenty of deployed policies and the labels of their performance are available. However, this assumption does not apply to most OPE scenarios because collecting such training data might be highly expensive. In this paper, we propose a novel OPR framework called SOCCER, where the existing OPE methods are modeled as workers in a crowdsourcing system. SOCCER can be trained in a self-supervised way as it does not require any ground-truth labels of policies. Moreover, in order to capture the relative discrepancies between policies, we propose a novel transformer-based architecture to learn effective pairwise policy representations. Experimental results show that SOCCER achieves significantly high accuracy in a variety of OPR tasks. Surprisingly, SOCCER even performs better than baselines trained in a supervised way using additional labeled data, which further demonstrates the superiority of SOCCER in OPR tasks.

****

Geometry Problem Solving based on Counterfactual Evolutionary Reasoning

SONG Bing,Xiong Gang,Fenghua Zhu,Lv Yisheng,Peijun Ye

As a representative topic in natural language processing and automated theorem proving, geometry problem solving requires an abstract problem understanding and symbolic reasoning. A major challenge here is to find a feasible reasoning sequence that is consistent with given axioms and the theorems already proved. Most recent methods have exploited neural network-based techniques to automatically di scover eligible solving steps. Such a kind of methods, however, is greatly impacted by the expert solutions for training. To improve the accuracy, this paper proposes a new method called counterfactual evolutionary reasoning, which uses a generative adversarial network to generate initial reasoning sequences and then introduces counterfactual reasoning to explore potential solutions. By directly exploring theorem candidates rather than the neural network selection, the new method can sufficiently extend the searching space to get a more appropriate reasoning step. Through comparative experiments on the recent proposed geometry3k, the largest geometry problem solving dataset, our method generally achieves a higher accuracy than most previous methods, bringing an overall improvement about 4.4% compared with the transformer models.

****

Few-Shot Domain Adaptation For End-to-End Communication

Jayaram Raghuram,Yijing Zeng,Dolores Garcia,Rafael Ruiz,Somesh Jha,Joerg Widmer,Suman Banerjee

The problem of end-to-end learning of a communication system using an autoencoder -- consisting of an encoder, channel, and decoder modeled using neural networks -- has recently been shown to be an effective approach. A challenge faced in the practical adoption of this learning approach is that under changing channel conditions (e.g. a wireless link), it requires frequent retraining of the autoencoder in order to maintain a low decoding error rate. Since retraining is both time consuming and requires a large number of samples, it becomes impractical when the channel distribution is changing quickly. We propose to address this problem using a fast and sample-efficient (few-shot) domain adaptation method that does not change the encoder and decoder networks. Different from conventional training-time unsupervised or semi-supervised domain adaptation, here we have a trained autoencoder from a source distribution that we want to adapt (at test time) to a target distribution using only a small labeled dataset, and no unlabeled data. We focus on a generative channel model based on the Gaussian mixture density network (MDN), and propose a regularized, parameter-efficient adaptation of the MDN using a set of affine transformations. The learned affine transformations are then used to design an optimal transformation at the decoder input to compensate for the distribution shift, and effectively present to the decoder inputs close to the source distribution. Experiments on many simulated distribution changes common to the wireless setting, and a real mmWave FPGA testbed demonstrate the effectiveness of our method at adaptation using very few target domain samples~\footnote{Code for our work: \url{https://github.com/jayaram-r/domain-adaptation-autoencoder}}.

****

HyPHEN: A Hybrid Packing Method and Optimizations for Homomorphic Encryption-Based Neural Network

Jaiyoung Park,Donghwan Kim,Jung Ho Ahn

Private Inference (PI) enables users to enjoy secure AI inference services while companies comply with regulations. Fully Homomorphic Encryption (FHE) based Convolutional Neural Network (CNN) inference is promising as users can offload the whole computation process to the server while protecting the privacy of sensitive data. Recent advances in AI research have enabled HE-friendly deep CNN like ResNet. However, FHE-based CNN (HCNN) suffers from high computational overhead.
Prior HCNN approaches rely on dense packing techniques that aggregate as many channels into the ciphertext to reduce element-wise operations like multiplication and bootstrapping.
However, these approaches require performing an excessive amount of homomorphic rotations to accumulate channels and maintain dense data organization, which takes up most of the runtime.
To overcome this limitation, we present HyPHEN, a deep HCNN implementation that drastically reduces the number of homomorphic rotations.
HyPHEN utilizes a novel convolution algorithm, RAConv, utilizing replication-based data organization, which leads to a significant reduction in rotation count.
Furthermore, we propose hybrid gap packing method for HyPHEN, which gathers sparse convolution results into a dense data organization with a marginal increase i

n the number of rotations.
HyPHEN explores the trade-off between the computational costs of rotations and other operations, and finds the optimal point minimizing the execution time.
With these optimizations, HyPHEN takes 3.8-4.9$\times$ less execution time than the state-of-the-art HCNN implementation and brings the runtimes of ResNet inference down to 1.38-14.86s using a GPU-accelerated HEAAN library.

****

Causal Inference for Knowledge Graph Completion

Changyi Xiao,Xiangnan He,Yixin Cao

The basis of existing knowledge graph completion (KGC) models is to learn the correlations in data, such as the correlation between entities or relations and scores of triplets. Since the world is driven by causality rather than correlation, correlation-driven KGC models are weak in interpretation and suffer from the data bias issue. In this paper, we propose causal KGC models to alleviate the issues by leveraging causal inference framework. Our models are intuitive and interpretable by utilizing causal graphs, controllable by using intervention techniques and model-agnostic. Causal graphs allow us to explain the causal relationships between variables and the data generation process. Under the causal graph, data bias can be seen as confounders. Then we block the bad effect of confounders by intervention operators to mitigate the data bias issue. Due to the difficulty of obtaining randomized data, causal KGC models pose unique challenges for evaluation. Thus, we show a method that makes evaluation feasible. Finally, we show a group theory view for KGC, which is equivalent to the view of causal but further reveals the relationships between causal graphs. Experimental results show that our causal KGC models achieve better performance than traditional KGC models.

****

Formal Specifications from Natural Language

Christopher Hahn,Frederik Schmitt,Julia Janice Tillman,Niklas Metzger,Julian Siber,Bernd Finkbeiner

We study the generalization abilities of language models when translating natural language into formal specifications with complex semantics. In particular, we fine-tune language models on three datasets consisting of English sentences and their corresponding formal representation: 1) regular expressions (regex), frequently used in programming and search; 2) First-order logic (FOL), commonly used in software verification and theorem proving; and 3) linear-time temporal logic (LTL), which forms the basis for industrial hardware specification languages. Our experiments show that, in these diverse domains, the language models maintain their generalization capabilities from pre-trained knowledge of natural language to generalize, e.g., to new variable names or operator descriptions. Additionally, they achieve competitive performance, and even outperform the state-of-the-art for translating into regular expressions, with the benefits of being easy to access, efficient to fine-tune, and without a particular need for domain-specific reasoning.

****

DELTA: Diverse Client Sampling for Fasting Federated Learning

Lin Wang,Yongxin Guo,Tao Lin,Xiaoying Tang

Partial client participation has been widely adopted in Federated Learning (FL) to efficiently reduce the communication burden. However, an improper client sampling scheme will select unrepresentative subsets, which will cause a large variance in the model update and slows down the convergence. Existing sampling methods are either biased or can be further improved to accelerate the convergence. In this paper, we propose an unbiased sampling scheme, termed DELTA, to alleviate this problem. In particular, DELTA characterizes the impact of client diversity and local variance and samples the representative clients who carry valuable information for global model updates. Moreover, DELTA is a provably optimal unbiased sampling scheme that minimizes the variance caused by partial client participation and achieves better convergence than other unbiased sampling schemes. We corroborate our results with experiments on both synthetic and real data sets.

****

Incremental Predictive Coding: A Parallel and Fully Automatic Learning Algorithm

Tommaso Salvatori,Yuhang Song,Beren Millidge,Zhenghua Xu,Lei Sha,Cornelius Emde,Rafal Bogacz,Thomas Lukasiewicz

Neuroscience-inspired models, such as predictive coding, have the potential to play an important role in the future of machine intelligence. However, they are not yet used in industrial applications due to some limitations, such as efficiency. In this work, we propose incremental predictive coding (iPC), a variation of the original model derived from the incremental expectation maximization algorithm, where every operation can be performed in parallel without external control. We show both theoretically and empirically that iPC is more efficient than the original algorithm by Rao and Ballard, with performances comparable to those of backpropagation in image classification tasks. This work impacts several areas, as it has general applications in computational neuroscience and machine learning, and specific applications in scenarios where automatization and parallelization are important, such as distributed computing and implementations of deep learning models on analog and neuromorphic chips.

****

Learning Geometric Representations of Interactive Objects

Alfredo Reichlin,Giovanni Luca Marchetti,Hang Yin,Anastasia Varava,Danica Kragic

We address the problem of learning geometric representations from observations perceived by an agent operating within an environment and interacting with an external object. To this end, we propose a representation learning framework that extracts the state of both the agent and the object from unstructured observations of arbitrary nature (e.g., images). Supervision comes from the performed actions alone, while the dynamics of the object is assumed to be unknown. We provide a theoretical foundation and formally prove that an ideal learner is guaranteed to infer an isometric representation, disentangling the agent from the object. Finally, we investigate empirically our framework on a variety of scenarios. Results show that our model reliably infers the correct representation and outperforms vision-based approaches such as a state-of-the-art keypoint extractor.

****

Improve distance metric learning by learning positions of class centers

Kun Song,Lantian Chu,Junwei Han,Fakhri Karray

Deep metric learning aims at learning a deep neural network by letting similar samples have small distances while dissimilar samples have large distances. To achieve this goal, the current DML algorithms mainly focus on pulling similar samples in each class as closely as possible. However, pulling similar samples only considers the local distribution of the data samples and ignores the global distribution of the data set, i.e., the center positions of different classes. The global distribution helps the distance metric learning. For example, expanding the distance between centers can increase the discriminant ability of the extracted features. However, how to increase the distance between centers is a challenging task. In this paper, we design a genius function named the skewed mean function, which only considers the most considerable distances of a set of samples. So maximizing the value of the skewed mean function can make the largest distance larger. We also prove that the current energy functions used for uniformity regularization on centers are special cases of our skewed mean function. At last, we conduct extensive experiments to illustrate the superiority of our methods.

****

The guide and the explorer: smart agents for resource-limited iterated batch reinforcement learning

Othman Gaizi,Albert Thomas,Balázs Kégl,Gabriel Hurtado

Iterated (a.k.a growing) batch reinforcement learning (RL) is a growing subfield fueled by the demand from systems engineers for intelligent control solutions that they can apply within their technical and organizational constraints. Model-based RL (MBRL) suits this scenario well for its sample efficiency and modularity. Recent MBRL techniques combine efficient neural system models with classical planning (like model predictive control; MPC). In this paper we add two components to this classical setup. The first is a Dyna-style policy learned on the system model using model-free techniques. We call it the guide since it guides the p lanner. The second component is the explorer, a strategy to expand the limited knowledge of the guide during planning. Through a rigorous ablation study we show that combination of these two ingredients is crucial for optimal performance and better data efficiency. We apply this approach with an off-policy guide and a heating explorer to improve the state of the art of benchmark systems addressing both discrete and continuous action spaces.

****

FairGBM: Gradient Boosting with Fairness Constraints
André Cruz,Catarina G Belém,João Bravo,Pedro Saleiro,Pedro Bizarro

Tabular data is prevalent in many high-stakes domains, such as financial services or public policy. Gradient Boosted Decision Trees (GBDT) are popular in these settings due to their scalability, performance, and low training cost. While fairness in these domains is a foremost concern, existing in-processing Fair ML methods are either incompatible with GBDT, or incur in significant performance losses while taking considerably longer to train. We present FairGBM, a dual ascent learning framework for training GBDT under fairness constraints, with little to no impact on predictive performance when compared to unconstrained GBDT. Since observational fairness metrics are non-differentiable, we propose smooth convex error rate proxies for common fairness criteria, enabling gradient-based optimization using a ``proxy-Lagrangian'' formulation. Our implementation shows an order of magnitude speedup in training time relative to related work, a pivotal aspect to foster the widespread adoption of FairGBM by real-world practitioners.

****

How (Un)Fair is Text Summarization?
Hannah Brown,Reza Shokri

Creating a good summary requires carefully choosing details from the original text to accurately represent it in a limited space. If a summary contains biased information about a group, it risks passing this bias off to readers as fact. These risks increase if we consider not just one biased summary, but rather a biased summarization algorithm. Despite this, little work has measured whether these summarizers demonstrate biased performance. Rather, most work in summarization focuses on improving performance, ignoring questions of bias. In this paper we demonstrate that automatic summarizers both amplify and introduce bias towards information about under-represented groups. Additionally, we show that summarizers are highly sensitive to document structure, making the summaries they generate unstable under changes that are semantically meaningless to humans, which poses a further fairness risk. Given these results, and the large scale potential for harm presented by biased summarization, we recommend that bias analysis be performed and reported on summarizers to ensure that new automatic summarization methods do not introduce bias to the summaries they generate.

****

Simulating Task-Free Continual Learning Streams From Existing Datasets
Aristotelis Chrysakis,Marie-Francine Moens

Task-free continual learning is the subfield of machine learning that focuses on learning online from a stream whose distribution changes continuously over time. However, previous works evaluate task-free continual learning using streams with distributions that change only at a few distinct points in time. In order to address the discrepancy between the definition and evaluation of task-free continual learning, we propose a principled algorithm that can permute any labeled dataset into a stream that is continuously nonstationary. We empirically show that the streams generated by our algorithm are less structured than the ones conventionally used in the literature. Moreover, we use our simulated task-free streams to benchmark multiple methods applicable to the task-free setting. We hope that our work will make it more likely that task-free continual learning methods are able to better generalize to real-world problems.

****

Online Bias Correction for Task-Free Continual Learning
Aristotelis Chrysakis,Marie-Francine Moens

Task-free continual learning is the machine-learning setting where a model is trained online with data generated by a nonstationary stream. Conventional wisdom suggests that, in this setting, models are trained using an approach called experience replay, where the risk is computed both with respect to current stream observations and to a small subset of past observations. In this work, we explain both theoretically and empirically how experience replay biases the outputs of the model towards recent stream observations. Moreover, we propose a simple approach to mitigate this bias online, by changing how the output layer of the model is optimized. We show that our approach improves significantly the learning performance of experience-replay approaches over different datasets. Our findings suggest that, when performing experience replay, the output layer of the model should be optimized separately from the preceding layers.

****

A Simple Contrastive Learning Objective for Alleviating Neural Text Degeneration

Shaojie Jiang,Ruqing Zhang,Svitlana Vakulenko,Maarten de Rijke

The cross-entropy objective has proved to be an all-purpose training objective for autoregressive language models (LMs). However, without distinguishing problematic tokens, LMs trained using cross-entropy exhibit text degeneration problems. To address this, unlikelihood training has been proposed to reduce the probability of unlikely tokens predicted by LMs. But unlikelihood does not explicitly consider the relationship between the label tokens and unlikely token candidates, thus showing marginal improvements in degeneration. We propose a new contrastive token learning objective that inherits the advantages of cross-entropy and unlikelihood training and avoids their limitations. The key idea is to teach a LM to generate high probabilities for label tokens and low probabilities for negative candidates. Comprehensive experiments on language modeling and open-domain dialogue generation tasks show that the proposed contrastive token objective yields much less repetitive texts, with a higher generation quality than baseline approaches, achieving the new state-of-the-art performance on text degeneration.

****

Enriching Online Knowledge Distillation with Specialist Ensemble

Mincheol Park,Woojeong Kim,Junsik Bang,Won Woo Ro,Suhyun Kim

Online Knowledge Distillation (KD) has an advantage over traditional KD works in that it removes the necessity for a pre-trained teacher. Indeed, an ensemble of small teachers has become typical guidance for a student's learning trajectory. Previous works emphasized diversity to create helpful ensemble knowledge and further argued that the size of diversity should be significant to prevent homogenization. This paper proposes a well-founded online KD framework with naturally derived specialists. In supervised learning, the parameters of a classifier are optimized by stochastic gradient descent based on a training dataset distribution. If the training dataset is shifted, the optimal point and corresponding parameters change accordingly, which is natural and explicit.

We first introduce a label prior shift to induce evident diversity among the same teachers, which assigns a skewed label distribution to each teacher and simultaneously specializes them through importance sampling. Compared to previous works, our specialization achieves the highest level of diversity and maintains it throughout training. Second, we propose a new aggregation that uses post-compensation in specialist outputs and conventional model averaging. The aggregation empirically exhibits the advantage of ensemble calibration even if applied to previous diversity-eliciting methods. Finally, through extensive experiments, we demonstrate the efficacy of our framework on top-1 error rate, negative log-likelihood, and notably expected calibration error.

****

Improved Gradient Descent Optimization Algorithm based on Inverse Model-Parameter Difference

Ayushya Pare,Zhichun Lei

A majority of deep learning models implement first-order optimization algorithms like the stochastic gradient descent (SGD) or its adaptive variants for training large neural networks. However, slow convergence due to complicated geometry of the loss function is one of the major challenges faced by the SGD. The currently popular optimization algorithms incorporate an accumulation of past gradients to improve the gradient descent convergence via either the accelerated gradient

scheme (including Momentum, NAG, etc.) or the adaptive learning-rate scheme (including Adam, AdaGrad, etc.). Despite their general popularity, these algorithms often display suboptimal convergence owing to extreme scaling of the learning-rate due to the accumulation of past gradients. In this paper, a novel approach to gradient descent optimization is proposed which utilizes the difference in the model-parameter values from the preceding iterations to adjust the learning-rate of the algorithm. More specifically, the learning-rate for each model-parameter is adapted inversely proportional to the displacement of the model-parameter from the previous iterations. As the algorithm utilizes the displacement of model-parameters, poor convergence caused due to the accumulation of past gradients is avoided. A convergence analysis based on the regret bound approach is performed and the theoretical bounds for a stable convergence are determined. An Empirical analysis evaluates the proposed algorithm applied on the CIFAR 10/100 and the ImageNet datasets and compares it with the currently popular optimizers. The experimental results demonstrate that the proposed algorithm shows better performance than the popular optimization algorithms.
****
Variational Learning ISTA
Fabio Valerio Massoli,Christos Louizos,Arash Behboodi
Compressed sensing combines the power of convex optimization techniques with a sparsity inducing prior on the signal space to solve an underdetermined system of equations. For many problems, the sparsifying dictionary is not directly given, nor its existence can be assumed. Besides, the sensing matrix can change across different scenarios. Addressing these issues requires solving a sparse representation learning problem, namely dictionary learning, taking into account the epistemic uncertainty on the learned dictionaries and, finally, jointly learning sparse representations and reconstructions under varying sensing matrix conditions.

We propose a variant of the LISTA architecture that incorporates the sensing matrix into the architecture. In particular, we propose to learn a distribution over dictionaries via a variational approach, dubbed \ac{VLISTA}, which approximates a posterior distribution over the dictionaries as part of an unfolded LISTA-based recovery network. Such a variational posterior distribution is updated after each iteration, and thereby adapts the dictionary according to the optimization dynamics. As a result, \ac{VLISTA} provides a probabilistic way to jointly learn the dictionary distribution and the reconstruction algorithm with varying sensing matrices. We provide theoretical and experimental support for our architecture and show that it learns calibrated uncertainties.
****
Moment Distributionally Robust Probabilistic Supervised Learning
Yeshu Li,Brian D Ziebart
Probabilistic supervised learning assumes the groundtruth itself is a distribution instead of a single label, as in classic settings. Common approaches learn with a proper composite loss and obtain probability estimates via an invertible link function. Typical links such as the softmax yield restrictive and problematic uncertainty certificates. In this paper, we propose to make direct prediction of conditional label distributions from first principles in distributionally robust optimization based on an ambiguity set defined by feature moment divergence. We derive its generalization bounds under mild assumptions. We illustrate how to manipulate penalties for underestimation and overestimation. Our method can be easily incorporated into neural networks for end-to-end representation learning. Experimental results on datasets with probabilistic labels illustrate the flexibility, effectiveness, and efficiency of this learning paradigm.
****
CLEP: Exploiting Edge Partitioning for Graph Contrastive Learning
Yilin He,Chaojie Wang,Hao Zhang,Bo Chen,Bo An,Mingyuan Zhou
Generative and contrastive are two fundamental unsupervised approaches to model graph information. The graph generative models extract intra-graph information whereas the graph contrastive learning methods focus on inter-graph information. Combining these complementary sources of information can potentially enhance the

expressiveness of graph representations, which, nevertheless, is underinvestigated by existing methods. In this work, we introduce a probabilistic framework called contrastive learning with edge partitioning (CLEP) that integrates generative modeling and graph contrastive learning. CLEP models edge generation by cumulative latent node interactions over multiple mutually independent hidden communities. Inspired by the ``assembly'' behavior of communities in graph generation, CEGCL learns community-specific graph embeddings and assemble them together to represent the entire graph, which are further used to predict the graph's identity via a contrastive objective. To relate each embedding to one hidden community, we define a set of community-specific weighted edges for node feature aggregation by partitioning the observed edges according to the latent node interactions associated with the corresponding hidden community. With these unique designs, CLEP is able to model the statistical dependency among hidden communities, graph structures as well as the identity of each graph; it can also be trained end-to-end via variational inference. We evaluate CLEP on real-world benchmarks under self-supervised and semi-supervised settings and achieve promising results, which demostrate the effectiveness of our method. Various exploratory studies are also conducted to highlight the characteristics of the inferred hidden communities and the potential benefits they bring to representation learning.

****

Meta-Learning the Inductive Biases of Simple Neural Circuits

Will Dorrell,Maria Yuffa,Peter E. Latham

Animals receive noisy and incomplete information, from which we must learn how to react in novel situations. A fundamental problem is that training data is always finite, making it unclear how to generalise to unseen data. But, animals do react appropriately to unseen data, wielding Occam's razor to select a parsimonious explanation of the observations. How they do this is called their inductive bias, and it is implicitly built into the operation of animals' neural circuits. This relationship between an observed circuit and its inductive bias is a useful explanatory window for neuroscience, allowing design choices to be understood normatively. However, it is generally very difficult to map circuit structure to inductive bias. In this work we present a neural network tool to bridge this gap. The tool allows us to meta-learn the inductive bias of neural circuits by learning functions that a neural circuit finds easy to generalise, since easy-to-generalise functions are exactly those the circuit chooses to explain incomplete data. We show that in systems where the inductive bias is known analytically, i.e. linear and kernel regression, our tool recovers it. Then, we show it is able to flexibly extract inductive biases from differentiable circuits, including spiking neural networks. This illustrates the intended use case of our tool: understanding the role of otherwise opaque pieces of neural functionality, such as non-linearities, learning rules, or connectomic data, through the inductive bias they induce.

****

Accelerating spiking neural network training using the d-block model

Luke Taylor,Andrew J King,Nicol Spencer Harper

There is a growing interest in using spiking neural networks (SNNs) to study the brain \textit{in silico} and in emulating them on neuromorphic computers due to their lower energy consumption compared to artificial neural networks (ANNs). Significant progress has been made in directly training SNNs to perform on a par with ANNs in terms of accuracy. However, these methods are slow due to their sequential nature and require careful network regularisation to avoid overfitting. We propose a new SNN model, the d-block model, with stochastic absolute refractory periods and recurrent conductance latencies, which reduces the number of sequential computations using fast vectorised operations. Our model obtains accelerated training speeds and state-of-the-art performance across various neuromorphic datasets without the need for any regularisation and using fewer spikes compared to standard SNNs.

****

RG: OUT-OF-DISTRIBUTION DETECTION WITH REACTIVATE GRADNORM

Mingyu Xu,Kexin Wang,Zheng Lian,Licai Sun,Bin Liu,Jianhua Tao

Detecting out-of-distribution (OOD) data is critical to building reliable machine learning systems in the open world. Previous works mainly perform OOD detection in feature space or output space. Recently, researchers have achieved promising results using gradient information, which combines the information in both feature and output space for OOD detection. However, existing works still suffer from the problem of overconfidence. To address this problem, we propose a novel method called ``Reactivate Gradnorm (RG)'', which exploits the norm of the clipped feature vector and the energy in the output space for OOD detection. To verify the effectiveness of our method, we conduct experiments on four benchmark datasets. Experimental results demonstrate that our RG outperforms existing state-of-the-art approaches by 2.06\% in average AUROC. Meanwhile, RG is easy to implement and does not require additional OOD data or fine-tuning process. We can realize OOD detection in only one forward pass of any pretrained model.

****

Don’t fear the unlabelled: safe semi-supervised learning via debiasing

Hugo Schmutz,Olivier HUMBERT,Pierre-Alexandre Mattei

Semi-supervised learning (SSL) provides an effective means of leveraging unlabelled data to improve a model’s performance. Even though the domain has received a considerable amount of attention in the past years, most methods present the common drawback of lacking theoretical guarantees. Our starting point is to notice that the estimate of the risk that most discriminative SSL methods minimise is biased, even asymptotically. This bias impedes the use of standard statistical learning theory and can hurt empirical performance. We propose a simple way of removing the bias. Our debiasing approach is straightforward to implement and applicable to most deep SSL methods. We provide simple theoretical guarantees on the trustworthiness of these modified methods, without having to rely on the strong assumptions on the data distribution that SSL theory usually requires. In particular, we provide generalisation error bounds for the proposed methods. We evaluate debiased versions of different existing SSL methods, such as the Pseudo-label method and Fixmatch, and show that debiasing can compete with classic deep SSL techniques in various settings by providing better calibrated models. Additionally, we provide a theoretical explanation of the intuition of the popular SSL methods. An implementation of a debiased version of Fixmatch is available at https://github.com/HugoSchmutz/DeFixmatch

****

Gandalf : Data Augmentation is all you need for Extreme Classification

Siddhant Kharbanda,Devaansh Gupta,Erik Schultheis,Atmadeep Banerjee,Vikas Verma,Rohit Babbar

Extreme Multi-label Text Classification (XMC) involves learning a classifier that can assign an input with a subset of most relevant labels from millions of label choices. Recent works in this domain have increasingly focused on the problem setting with short-text input data, and labels endowed with short textual descriptions called label features. Short-text XMC with label features has found numerous applications in areas such as prediction of related searches, title-based product recommendation, bid-phrase suggestion, amongst others. In this paper, we propose Gandalf, a graph induced data augmentation based on label features, such that the generated data-points can supplement the training distribution. By exploiting the characteristics of the short-text XMC problem, it leverages the label features to construct valid training instances, and uses the label graph for generating the corresponding soft-label targets, hence effectively capturing the label-label correlations. While most recent advances (such as SiameseXML and ECLARE) in XMC have been algorithmic, mainly aimed towards developing novel deep-learning architectures, our data-centric augmentation approach is orthogonal to these methodologies. We demonstrate the generality and effectiveness of Gandalf by showing up to 30% relative improvements for 5 state-of-the-art algorithms across 4 benchmark datasets consisting of up to 1.3 million labels.

****

Learning a Data-Driven Policy Network for Pre-Training Automated Feature Engineering

Liyao Li,Haobo Wang,Liangyu Zha,Qingyi Huang,Sai Wu,Gang Chen,Junbo Zhao

Feature engineering is widely acknowledged to be pivotal in tabular data analysis and prediction. Automated feature engineering (AutoFE) emerged to automate this process managed by experienced data scientists and engineers conventionally. In this area, most — if not all — prior work adopted an identical framework from the neural architecture search (NAS) method. While feasible, we posit that the NAS framework very much contradicts the way how human experts cope with the data since the inherent Markov decision process (MDP) setup differs. We point out that its data-unobserved setup consequentially results in an incapability to generalize across different datasets as well as also high computational cost. This paper proposes a novel AutoFE framework Feature Set Data-Driven Search (FETCH), a pipeline mainly for feature generation and selection. Notably, FETCH is built on a brand-new data-driven MDP setup using the tabular dataset as the state fed into the policy network. Further, we posit that the crucial merit of FETCH is its transferability where the yielded policy network trained on a variety of datasets is indeed capable to enact feature engineering on unseen data, without requiring additional exploration. To the best of our knowledge, this is a pioneer attempt to build a tabular data pre-training paradigm via AutoFE. Extensive experiments show that FETCH systematically surpasses the current state-of-the-art AutoFE methods and validates the transferability of AutoFE pre-training.

****

Attention Flows for General Transformers
Niklas Metzger,Christopher Hahn,Julian Siber,Frederik Schmitt,Bernd Finkbeiner

In this paper, we study the computation of how much an input token in a Transformer model influences its prediction. We formalize a method to construct a flow network out of the attention values of encoder-only Transformer models and extend it to general Transformer architectures, including an auto-regressive decoder. We show that running a maxflow algorithm on the flow network construction yields Shapley values, which determine a player's impact in cooperative game theory. By interpreting the input tokens in the flow network as players, we can compute their influence on the total attention flow leading to the decoder's decision. Additionally, we provide a library that computes and visualizes the attention flow of arbitrary Transformer models. We show the usefulness of our implementation on various models trained on natural language processing and reasoning tasks.

****

Making Substitute Models More Bayesian Can Enhance Transferability of Adversarial Examples
Qizhang Li,Yiwen Guo,Wangmeng Zuo,Hao Chen

The transferability of adversarial examples across deep neural networks (DNNs) is the crux of many black-box attacks. Many prior efforts have been devoted to improving the transferability via increasing the diversity in inputs of some substitute models. In this paper, by contrast, we opt for the diversity in substitute models and advocate to attack a Bayesian model for achieving desirable transferability. Deriving from the Bayesian formulation, we develop a principled strategy for possible finetuning, which can be combined with many off-the-shelf Gaussian posterior approximations over DNN parameters. Extensive experiments have been conducted to verify the effectiveness of our method, on common benchmark datasets, and the results demonstrate that our method outperforms recent state-of-the-arts by large margins (roughly 19% absolute increase in average attack success rate on ImageNet), and, by combining with these recent methods, further performance gain can be obtained. Our code: https://github.com/qizhangli/MoreBayesian-attack.

****

Learning Group Importance using the Differentiable Hypergeometric Distribution
Thomas M. Sutter,Laura Manduchi,Alain Ryser,Julia E Vogt

Partitioning a set of elements into subsets of a priori unknown sizes is essential in many applications. These subset sizes are rarely explicitly learned - be it the cluster sizes in clustering applications or the number of shared versus independent generative latent factors in weakly-supervised learning. Probability distributions over correct combinations of subset sizes are non-differentiable due to hard constraints, which prohibit gradient-based optimization. In this work, we propose the differentiable hypergeometric distribution. The hypergeometric distribution models the probability of different group sizes based on their relative importance. We introduce reparameterizable gradients to learn the importance between groups and highlight the advantage of explicitly learning the size of subsets in two typical applications: weakly-supervised learning and clustering. In both applications, we outperform previous approaches, which rely on suboptimal heuristics to model the unknown size of groups.

****

Convergence Rate of Primal-Dual Approach to Constrained Reinforcement Learning with Softmax Policy

Long Yang,Li Shen,Pengfei Li,Yaodong Yang,Zhouchen Lin,Gang Pan

In this paper, we consider primal-dual approach to solve constrained reinforcement learning (RL) problems, where we formulate constrained reinforcement learning under constrained Markov decision process (CMDP). We propose the primal-dual policy gradient (PD-PG) algorithm with softmax policy. Although the constrained RL involves a non-concave maximization problem over the policy parameter space, we show that for both exact policy gradient and model-free learning, the proposed PD-PG needs iteration complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ to achieve its optimal policy for both constraint and reward performance. Such an iteration complexity outperforms or matches most constrained RL algorithms. For the learning with exact policy gradient, the main challenge is to show the positivity of deterministic optimal policy (at the optimal action) is independent on both state space and iteration times.

For the model-free learning, since we consider the discounted infinite-horizon setting, and the simulator can not rollout with an infinite-horizon sequence; thus one of the main challenges lies in how to design unbiased value function estimators with finite-horizon trajectories. We consider the unbiased estimators with finite-horizon trajectories that involve geometric distribution horizons, which is the key technique for us to obtain the theoretical results for model-free learning.

****

Cross-Layer Retrospective Retrieving via Layer Attention

Yanwen Fang,Yuxi CAI,Jintai Chen,Jingyu Zhao,Guangjian Tian,Guodong Li

More and more evidence has shown that strengthening layer interactions can enhance the representation power of a deep neural network, while self-attention excels at learning interdependencies by retrieving query-activated information. Motivated by this, we devise a cross-layer attention mechanism, called multi-head recurrent layer attention (MRLA), that sends a query representation of the current layer to all previous layers to retrieve query-related information from different levels of receptive fields. A light-weighted version of MRLA is also proposed to reduce the quadratic computation cost. The proposed layer attention mechanism can enrich the representation power of many state-of-the-art vision networks, including CNNs and vision transformers. Its effectiveness has been extensively evaluated in image classification, object detection and instance segmentation tasks, where improvements can be consistently observed. For example, our MRLA can improve 1.6% Top-1 accuracy on ResNet-50, while only introducing 0.16M parameters and 0.07B FLOPs. Surprisingly, it can boost the performances by a large margin of 3-4% box AP and mask AP in dense prediction tasks. Our code is available at https://github.com/joyfang1106/MRLA.

****

Decision S4: Efficient Sequence-Based RL via State Spaces Layers

Shmuel Bar David,Itamar Zimerman,Eliya Nachmani,Lior Wolf

Recently, sequence learning methods have been applied to the problem of off-policy
Reinforcement Learning, including the seminal work on Decision Transformers,
which employs transformers for this task. Since transformers are parameter-heavy,
cannot benefit from history longer than a fixed window size, and are not computed
using recurrence, we set out to investigate the suitability of the S4 family of

models, which are based on state-space layers and have been shown to outperform transformers, especially in modeling long-range dependencies. In this work, we present two main algorithms: (i) an off-policy training procedure that works with trajectories, while still maintaining the training efficiency of the S4 model. (ii) An on-policy training procedure that is trained in a recurrent manner, benefits from long-range dependencies, and is based on a novel stable actor-critic mechanism. Our results indicate that our method outperforms multiple variants of decision transformers, as well as the other baseline methods on most tasks, while reducing the latency, number of parameters, and training time by several orders of magnitude, making our approach more suitable for real-world RL

****

Deep autoregressive density nets vs neural ensembles for model-based offline reinforcement learning

Abdelhakim Benechehab,Albert Thomas,Balázs Kégl

We consider the problem of offline reinforcement learning where only a set of system transitions is made available for policy optimization. Following recent advances in the field, we consider a model-based reinforcement learning algorithm that infers the system dynamics from the available data and performs policy optimization on imaginary model rollouts. This approach is vulnerable to exploiting model errors which can lead to catastrophic failures on the real system. The standard solution is to rely on ensembles for uncertainty heuristics and to avoid exploiting the model where it is too uncertain. We challenge the popular belief that we must resort to ensembles by showing that better performance can be obtained with a single well-calibrated autoregressive model on the D4RL benchmark. We also analyze static metrics of model-learning and conclude on the important model properties for the final performance of the agent.

****

Unveiling the sampling density in non-uniform geometric graphs

Raffaele Paolino,Aleksandar Bojchevski,Stephan Günnemann,Gitta Kutyniok,Ron Levie

A powerful framework for studying graphs is to consider them as geometric graphs: nodes are randomly sampled from an underlying metric space, and any pair of nodes is connected if their distance is less than a specified neighborhood radius. Currently, the literature mostly focuses on uniform sampling and constant neighborhood radius. However, real-world graphs are likely to be better represented by a model in which the sampling density and the neighborhood radius can both vary over the latent space. For instance, in a social network communities can be modeled as densely sampled areas, and hubs as nodes with larger neighborhood radius. In this work, we first perform a rigorous mathematical analysis of this (more general) class of models, including derivations of the resulting graph shift operators. The key insight is that graph shift operators should be corrected in order to avoid potential distortions introduced by the non-uniform sampling. Then, we develop methods to estimate the unknown sampling density in a self-supervised fashion. Finally, we present exemplary applications in which the learnt density is used to 1) correct the graph shift operator and improve performance on a variety of tasks, 2) improve pooling, and 3) extract knowledge from networks. Our experimental findings support our theory and provide strong evidence for our model.

****

Boosting Causal Discovery via Adaptive Sample Reweighting

An Zhang,Fangfu Liu,Wenchang Ma,Zhibo Cai,Xiang Wang,Tat-Seng Chua

Under stringent model type and variable distribution assumptions, score-based causal discovery methods learn the directed acyclic graph (DAG) from observational data by evaluating candidate graphs over an averaged score function. Despite the great success in low-dimensional linear systems, it has been observed that the se approaches overly exploits easier-to-fit samples, thus inevitably learning spurious edges. Worse still, the common homogeneity assumption of most causal discovery methods can be easily violated due to the widespread existence of heterogeneous data in the real world, resulting in performance vulnerability when noise distributions vary. We propose a simple yet effective model-agnostic framework to boost causal discovery performance by dynamically learning the adaptive weights for the Reweighted Score function, ReScore for short, where the learned weights tailors quantitatively to the important degree of each samples. Intuitively, we leverage the bilevel optimization scheme to alternatively train a standard DAG learner first, then upweight the samples that the DAG learner fails to fit well and downweight the samples that the DAG learner easily extracts the causation information from. Extensive experiments on both synthetic and real-world datasets are carried out to validate the effectiveness of ReScore. We observe consistent and significant boosts in structure learning performance. We further visualize that ReScore concurrently mitigates the influence of spurious edges and generalizes to heterogeneous data. Finally, we perform theoretical analysis to guarantee the structure identifiability and the weight adaptive properties of ReScore. Our codes are available at https://github.com/anzhang314/ReScore.

****

Robust Training through Adversarially Selected Data Subsets

Hitvarth Diwanji,Divyanshu Shende,Rishi Agarwal,Swaprava Nath,Abir De

Robustness to adversarial perturbations often comes at the cost of a drop in accuracy on unperturbed or clean instances. Most existing defense mechanisms attempt to defend the learner from attack on all possible instances, which often degrades the accuracy on clean instances significantly. However, in practice, an attacker might only select a small subset of instances to attack, $e.g.$, in facial recognition systems an adversary might aim to target specific faces. Moreover, the subset selection strategy of the attacker is seldom known to the defense mechanism a priori, making it challenging to attune the mechanism beforehand. This motivates designing defense mechanisms which can (i) defend against attacks on subsets instead of all instances to prevent degradation of clean accuracy and, (ii) ensure good overall performance for attacks on any selected subset. In this work, we take a step towards solving this problem. We cast the training problem as a min-max game involving worst-case subset selection along with optimization of model parameters, rendering the problem NP-hard. To tackle this, we first show that, for a given learner's model, the objective can be expressed as a difference between a γ-weakly submodular and a modular function. We use this property to propose ROGET, an iterative algorithm, which admits approximation guarantees for a class of loss functions. Our experiments show that ROGET obtains better overall accuracy compared to several state-of-the-art defense methods for different adversarial subset selection techniques.

****

Beyond Reward: Offline Preference-guided Policy Optimization

Yachen Kang,Diyuan Shi,Jinxin Liu,Li He,Donglin Wang

In this work, we study offline preference-based reinforcement learning (PbRL), which relaxes the two fundamental supervisory signals in standard reinforcement learning (online accessible transition dynamics and rewards). In other words, the agent is provided with fixed offline trajectory transitions and human preferences between pairs of trajectories. Due to the orthogonality property of rewards and dynamics, one common practice is combining prior PbRL-based reward learning objectives with off-the-shelf offline RL algorithms to bridge preference modeling and offline learning. However, such two isolated optimizations require learning a separate reward function and thus place an information bottleneck on reward learning (the bridge). As an alternative, we propose offline preference-guided policy optimization (OPPO), an end-to-end offline PbRL formulation, which jointly learns to model the preference (for finding the optimal task policy) and the offline data (for eliminating OOD). In particular, OPPO introduces an offline hindsight information matching objective and a preference modeling objective. Then, iterating the two objectives over, we can directly extract a well-performing decision policy, avoiding a separate reward learning. We empirically show that OPPO can effectively model the offline preference and outperform prior competing baselines (including the offline RL algorithms performed over the true reward function).

****

Iterative Circuit Repair Against Formal Specifications

Matthias Cosler,Frederik Schmitt,Christopher Hahn,Bernd Finkbeiner

We present a deep learning approach for repairing sequential circuits against formal specifications given in linear-time temporal logic (LTL). Given a defective circuit and its formal specification, we train Transformer models to output circuits that satisfy the corresponding specification. We propose a separated hierarchical Transformer for multimodal representation learning of the formal specification and the circuit. We introduce a data generation algorithm that enables generalization to more complex specifications and out-of-distribution datasets. In addition, our proposed repair mechanism significantly improves the automated synthesis of circuits from LTL specifications with Transformers. It improves the state-of-the-art by 6.8 percentage points on held-out instances and 11.8 percentage points on an out-of-distribution dataset from the annual reactive synthesis competition.

****

Neural Probabilistic Logic Programming in Discrete-Continuous Domains

Lennert De Smet,Pedro Zuidberg Dos Martires,Robin Manhaeve,Giuseppe Marra,Angelika Kimmig,Luc De Raedt

Neural-symbolic AI (NeSy) methods allow neural networks to exploit symbolic background knowledge. NeSy has been shown to aid learning in the limited data regime and to facilitate inference on out-of-distribution data. Neural probabilistic logic programming (NPLP) is a popular NeSy approach that integrates probabilistic models with neural networks and logic programming. A major limitation of current NPLP systems, such as DeepProbLog, is their restriction to discrete and finite probability distributions, e.g., binary random variables. To overcome this limitation, we introduce DeepSeaProbLog, an NPLP language that supports discrete and continuous random variables on (possibly) infinite and even uncountable domains. Our main contributions are 1) the introduction of DeepSeaProbLog and its semantics, 2) an implementation of DeepSeaProbLog that supports inference and gradient-based learning, and 3) an experimental evaluation of our approach.

****

Can BERT Refrain from Forgetting on Sequential Tasks? A Probing Study

Mingxu Tao,Yansong Feng,Dongyan Zhao

Large pre-trained language models have helped to achieve state of the art on a variety of NLP tasks, nevertheless, they still suffer from forgetting when incrementally learning a series of sequential tasks. To alleviate this problem, recent works propose several models enhanced by sparse experience replay and local adaption, which yield satisfactory performance. However, in this paper we find that pre-trained language models like BERT have a potential ability to learn sequentially, even without any sparse memory replay. To verify the ability of BERT to maintain old knowledge, we adopt and re-finetune single-layer probe networks with the parameters of BERT fixed. We investigate the models on two typical kinds of NLP tasks, text classification and extractive question answering. And our experiments reveal that BERT can actually generate high quality representations for previous tasks in a long term, under extremely sparse replay or even no replay. We further introduce a series of methods to interpret the mechanism of forgetting and how memory rehearsal plays a significant role in task incremental learning, which bridges the gap between our new discovery and previous studies about catastrophic forgetting. Additionally, we provide both quantified and visualized results demonstrating that the representation space of BERT is always topologically organised, which guarantees its performance.

****

Behavior Proximal Policy Optimization

Zifeng Zhuang,Kun LEI,Jinxin Liu,Donglin Wang,Yilang Guo

Offline reinforcement learning (RL) is a challenging setting where existing off-policy actor-critic methods perform poorly due to overestimating of out-of-distr ibution state-action pairs. Thus, various additional augmentations are proposed to keep the learned policy close to the offline dataset (or the behavior policy). In this work, starting from the analysis of offline monotonic policy improvement, we reach a surprising conclusion that online on-policy algorithms are naturally able to solve offline RL. Specifically, the inherent conservatism of these on-policy algorithms is exactly what the offline RL method needs to overcome the overestimation. Based on this, we propose Behavior Proximal Policy Optimization (BPPO), which solves offline RL without any extra constraint or regularization introduced compared to PPO. Extensive experiments on the D4RL benchmark empirically show this extremely succinct method outperforms state-of-the-art offline RL algorithms. Our implementation is available at https://github.com/Dragon-Zhuang/BPPO.

****

FedGC: An Accurate and Efficient Federated Learning under Gradient Constraint for Heterogeneous Data

Peng Liu,Jie Du,Chi Man VONG

Federated Learning (FL) is an important paradigm in large-scale distributed machine learning, which enables multiple clients to jointly learn a unified global model without transmitting their local data to a central server. FL has attracted growing attentions in many real-world applications, such as multi-center cardiovascular disease diagnosis and autonomous driving. Practically, the data across clients are always heterogeneous, i.e., not independently and identically distributed (Non-IID), making the local models suffer from catastrophic forgetting of the initial (or global) model. To mitigate this forgetting issue, existing FL methods may require additional regularization terms or generates pseudo data, resulting to 1) limited accuracy; 2) long training time and slow convergence rate for real-time applications; and 3) high communication cost. In this work, an accurate and efficient Federated Learning algorithm under Gradient Constraints (FedGC) is proposed, which provides three advantages: i) High accuracy is achieved by the proposed Client-Gradient-Constraint based projection method (CGC) to alleviate the forgetting issue occurred in clients, and the proposed Server-Gradient-Constraint based projection method (SGC) to effectively aggregate the gradients of clients; ii) Short training time and fast convergence rate are enabled by the proposed fast Pseudo-gradient-based mini-batch Gradient Descent (PGD) method and SGC; iii) Low communication cost is required due to the fast convergence rate and only gradients are necessary to be transmitted between server and clients. In the experiments, four real-world image datasets with three Non-IID types are evaluated, and five popular FL methods are used for comparison. The experimental results demonstrate that our FedGC not only significantly improves the accuracy and convergence rate on Non-IID data, but also drastically decreases the training time. Compared to the state-of-art FedReg, our FedGC improves the accuracy by up to 14.28% and speeds up the local training time by 15.5 times while decreasing 23% of the communication cost.

****

Actionable Neural Representations: Grid Cells from Minimal Constraints

Will Dorrell,Peter E. Latham,Timothy E. J. Behrens,James C. R. Whittington

To afford flexible behaviour, the brain must build internal representations that mirror the structure of variables in the external world. For example, 2D space obeys rules: the same set of actions combine in the same way everywhere (step north, then south, and you won't have moved, wherever you start). We suggest the brain must represent this consistent meaning of actions across space, as it allows you to find new short-cuts and navigate in unfamiliar settings. We term this representation an `actionable representation'. We formulate actionable representations using group and representation theory, and show that, when combined with biological and functional constraints - non-negative firing, bounded neural activity, and precise coding - multiple modules of hexagonal grid cells are the optimal representation of 2D space. We support this claim with intuition, analytic justification, and simulations. Our analytic results normatively explain a set of surprising grid cell phenomena, and make testable predictions for future experiments. Lastly, we highlight the generality of our approach beyond just understan ding 2D space. Our work characterises a new principle for understanding and designing flexible internal representations: they should be actionable, allowing animals and machines to predict the consequences of their actions, rather than just encode.

****

Compression-aware Training of Neural Networks using Frank-Wolfe

Max Zimmer,Christoph Spiegel,Sebastian Pokutta

Many existing Neural Network pruning approaches either rely on retraining to compensate for pruning-caused performance degradation or they induce strong biases to converge to a specific sparse solution throughout training. A third paradigm, ’compression-aware’ training, obtains state-of-the-art dense models which are robust to a wide range of compression ratios using a single dense training run while also avoiding retraining. In that vein, we propose a constrained optimization framework centered around a versatile family of norm constraints and the Stochastic Frank-Wolfe (SFW) algorithm which together encourage convergence to well-performing solutions while inducing robustness towards convolutional filter pruning and low-rank matrix decomposition. Comparing our novel approaches to compression methods in these domains on benchmark image-classification architectures and datasets, we find that our proposed scheme is able to yield competitive results, often outperforming existing compression-aware approaches. In the case of low-rank matrix decomposition, our approach can require much less computational resources than nuclear-norm regularization based approaches by requiring only a fraction of the singular values in each iteration. As a special case, our proposed constraints can be extended to include the unstructured sparsity-inducing constraint proposed constraint by Pokutta et al. (2020) and Miao et al. (2022), which we improve upon. Our findings also indicate that the robustness of SFW-trained models largely depends on the gradient rescaling of the learning rate and we establish a theoretical foundation for that practice.

****

Modeling content creator incentives on algorithm-curated platforms

Jiri Hron,Karl Krauth,Michael Jordan,Niki Kilbertus,Sarah Dean

Content creators compete for user attention. Their reach crucially depends on algorithmic choices made by developers on online platforms. To maximize exposure, many creators adapt strategically, as evidenced by examples like the sprawling search engine optimization industry. This begets competition for the finite user attention pool. We formalize these dynamics in what we call an exposure game, a model of incentives induced by modern algorithms including factorization and (deep) two-tower architectures. We prove that seemingly innocuous algorithmic choices—e.g., non-negative vs. unconstrained factorization—significantly affect the existence and character of (Nash) equilibria in exposure games. We proffer use of creator behavior models like ours for an (ex-ante) pre-deployment audit. Such an audit can identify misalignment between desirable and incentivized content, and thus complement post-hoc measures like content filtering and moderation. To this end, we propose tools for numerically finding equilibria in exposure games, and illustrate results of an audit on the MovieLens and LastFM datasets. Among else, we find that the strategically produced content exhibits strong dependence between algorithmic exploration and content diversity, and between model expressivity and bias towards gender-based user and creator groups.

****

MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals

Donghong Cai,Junru Chen,Yang Yang,Teng Liu,Yafeng Li

Brain signals are important quantitative data for understanding physiological activities and diseases of human brain. Meanwhile, rapidly developing deep learning methods offer a wide range of opportunities for better modeling brain signals, which has attracted considerable research efforts recently. Most existing studies pay attention to supervised learning methods, which, however, require high-cost clinical labels. In addition, the huge difference in the clinical patterns of brain signals measured by invasive (e.g., SEEG) and non-invasive (e.g., EEG) methods leads to the lack of a unified method. To handle the above issues, in this paper, we propose to study the self-supervised learning (SSL) framework for bra in signals that can be applied to pre-train either SEEG or EEG data. Intuitively, brain signals, generated by the firing of neurons, are transmitted among different connecting structures in human brain. Inspired by this, we propose to learn implicit spatial and temporal correlations between different channels (i.e., contacts of the electrode, corresponding to different brain areas) as the cornerstone for uniformly modeling different types of brain signals. Specifically, we capture the temporal correlation by designing the delayed-time-shift prediction task; we represent the spatial correlation by a graph structure, which is built with the goal to maximize the mutual information of each channel and its correlated ones. We further theoretically prove that our design can lead to a better predictive representation. Extensive experiments of seizure detection on both EEG and SEEG large-scale real- world datasets demonstrate our model outperforms several state-of-the-art time series SSL and unsupervised models.

****

Group-Disentangling Conditional Shift

Dan Andrei Iliescu,Damon Wischik

We propose a novel group disentanglement method called the Context-Aware Variational Autoencoder (CxVAE). Our model can learn disentangled representations on datasets with conditional shift. This phenomenon occurs when the conditional distribution of the instance-level latent variable $\mathbf{z}$ given the input observation $\mathbf{x}$ changes from one group to another (i.e. $p_i(\mathbf{z}|\mathbf{x}) \neq p_j(\mathbf{z}|\mathbf{x})$, where i,j are two different groups). We show that existing methods fail to learn disentangled representations under this scenario because they infer the group $\mathbf{u}$ and instance $\mathbf{z}$ variables separately. CxVAE overcomes this limitation by conditioning the instance inference on the group variable $q(\mathbf{z}|\mathbf{x},\mathbf{u})$. Our model has the novel ability to disentangle ambiguous observations (those with incomplete information about the generative factors), which we evaluate on the task of fair comparisons between student test scores. Additionally, we demonstrate empirically that conditional shift is the cause of our model's improved performance.

****

When and Why Is Pretraining Object-Centric Representations Good for Reinforcement Learning?

Jaesik Yoon,Yi-Fu Wu,Sungjin Ahn

Unsupervised object-centric representation (OCR) learning has recently been drawing a lot of attention as a new paradigm of visual representation. This is because of its potential of being an effective pretraining technique for various downstream tasks in terms of sample efficiency, systematic generalization, and reasoning. Although image-based reinforcement learning (RL) is one of the most important and thus frequently mentioned such downstream tasks, the benefit in RL has surprisingly not been investigated systematically thus far. Instead, most of the evaluations have focused on rather indirect metrics such as segmentation quality and object property prediction accuracy. In this paper, we investigate the effectiveness of OCR pretraining for image-based reinforcement learning via empirical experiments. For systematic evaluation, we introduce a simple object-centric visual RL benchmark and verify a series of hypotheses answering questions such as "Does OCR pretraining provide better sample efficiency?", "Which types of RL tasks benefit most from OCR pretraining?", and "Can OCR pretraining help with out-of-distribution generalization?". The results suggest that OCR pretraining is particularly effective in tasks where the relationship between objects is important, improving both task performance and sample efficiency when compared to single-vector representations. Furthermore, OCR models facilitate generalization to out-of-distribution tasks such as changing the number of objects or the appearance of the objects in the scene.

****

Face reconstruction from facial templates by learning latent space of a generator network

Hatef Otroshi Shahreza,Sébastien Marcel

Face recognition systems are increasingly deployed in different applications. In these systems, a feature vector (also called facial embeddings or templates) is typically extracted from each face image and is stored in the system's database during the enrollment stage, which is later used for comparison during the recognition stage. In this paper, we focus on the template inversion attack against face recognition systems and propose a new method to reconstruct face images from facial templates. Within a generative adversarial network (GAN)-based framework, we learn a mapping from facial templates to the intermediate latent space of a pre-trained face generation network, from which we can generate high-resolution realistic reconstructed face images. We show that our proposed method can be applied in whitebox and blackbox attacks against face recognition systems. Furthermore, we evaluate the transferability of our attack when the adversary uses the reconstructed face image to impersonate the underlying subject in an attack against another face recognition system. Considering the adversary's knowledge and the target face recognition system, we define five different attacks and evaluate the vulnerability of state-of-the-art face recognition systems. Our experiments show that our proposed method achieves high success attack rates in whitebox and blackbox scenarios. Furthermore, the reconstructed face images are transferable and can be used to enter target face recognition systems with a different feature extractor model.

****

Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules

Jun Xia,Chengshuai Zhao,Bozhen Hu,Zhangyang Gao,Cheng Tan,Yue Liu,Siyuan Li,Stan Z. Li

Recent years have witnessed the prosperity of pre-training graph neural networks (GNNs) for molecules. Typically, following the Masked Language Modeling (MLM) task of BERT~\citep{devlin2019bert}, \cite{hu2020strategies} first randomly mask the atom types and then pre-train the GNNs to predict them. However, unlike MLM, this pre-training task named AttrMask is too simple to learn informative molecular representations due to the extremely small and unbalanced atom vocabulary. As a remedy, we adopt the encoder of a variant of VQ-VAE~\citep{van2017neural} as a context-aware tokenizer to encode atoms as meaningful discrete values, which can enlarge the atom vocabulary size and mitigate the quantitative divergence between dominant (e.g., carbons) and rare atoms (e.g., phosphorus). With the enlarged atom vocabulary, we propose a novel node-level pre-training task, dubbed Masked Atoms Modeling (\textbf{MAM}), to randomly mask the discrete values and pre-train GNNs to predict them. MAM mitigates the negative transfer issue of AttrMask and can be combined with various pre-training tasks to advance their performance. Furthermore, for graph-level pre-training, we propose triplet masked contrastive learning (\textbf{TMCL}) to model varying degrees of semantic similarity between molecules, which is especially effective for molecule retrieval. MAM and TMCL constitute a novel pre-training framework, \textbf{Mole-BERT}, which can match or outperform state-of-the-art methods that require expensive domain knowledge as guidance. The codes, the tokenizer, and the pre-trained models will be released.

****

A sparse, fast, and stable representation for multiparameter topological data analysis

David Loiseaux,Mathieu Carrière,Andrew Blumberg

Topological data analysis (TDA) is a new area of geometric data analysis that focuses on using invariants from algebraic topology to provide multiscale shape descriptors for point clouds. One of the most important shape descriptors is persistent homology, which studies the topological variations as a filtration parameter changes; a typical parameter is the feature scale.

For many data sets, it is useful to consider varying multiple filtration parameters at once, for example scale and density. While the theoretical properties of one-parameter persistent homology are well understood, less is known about the multiparameter case. Of particular interest is the problem of representing multiparameter persistent homology by elements of a vector space for integration with traditional machine learning.

Existing approaches to this problem either ignore most of the multiparameter information to reduce to the one-parameter case or are heuristic and potentially unstable in the face of noise. In this article, we introduce a general representation framework for multiparameter persistent homology that encompasses previous approaches. We establish theoretical stability guarantees under this framework as well as efficient algorithms for practical computation, making this framework an applicable and versatile tool for TDA practitioners. We validate our stability results and algorithms with numerical experiments that demonstrate statistical convergence, prediction accuracy, and fast running times on several real data sets.

****

Improving Protein Interaction Prediction using Pretrained Structure Embedding

Chunchen Wang,YiWu Sun,Bing Yang,Shaochuan Li,Cheng Yang,Hui Li,Chuan Shi,Le Song

The prediction of protein-protein interactions (PPIs) is a critical problem because the knowledge of PPIs unravels the cellular behavior and its functionality. So far most previous works on PPI predictions mainly focused on sequence and network information and ignored the structural information of protein physical binding. We design a novel method, called xxx, which can leverage pretrained structure embedding and can be transferred to new ppi predictions. Experimental results on PPi predictions show that our pretrained structure embedding leads to significant improvement in PPI prediction comparing to sequence and network based methods. Furthermore, we show that embeddings pretrained based on ppi from different species can be transferred to improve the prediction for human proteins.

****

Batch Normalization and Bounded Activation Functions

Dongjin Kim,Woojeong Kim,TaeJoo Park,Suhyun Kim

Since Batch Normalization was proposed, it has been commonly located in front of activation functions, as proposed by the original paper. Swapping the order, i.e., using Batch Normalization after activation functions, has also been attempted, but it is generally not much different from the conventional order when ReLU is used. However, in the case of bounded activation functions like Tanh, we discovered that the swapped order achieves considerably better performance on various benchmarks and architectures than the conventional order. We report this remarkable phenomenon and closely examine what contributes to this performance improvement in this paper. One noteworthy thing about swapped models is the extreme saturation of activation values, which is usually considered harmful. Looking at the output distribution of individual activation functions, we found that many of them are highly asymmetrically saturated. The experiments inducing a different degree of asymmetric saturation support the hypothesis that asymmetric saturation helps improve performance. In addition, we found that Batch Normalization after bounded activation functions has another important effect: it relocates the asymmetrically saturated output of activation functions near zero. This enables the swapped model to have higher sparsity, further improving performance. Extensive experiments with Tanh, LeLecun Tanh, and Softsign show that the swapped models achieve improved performance with a high degree of asymmetric saturation.

****

MEDOE: A Multi-Expert Decoder and Output Ensemble Framework for Long-tailed Semantic Segmentation

Junao Shen,Long Chen,Tian Feng,Zhang Wei,Fei Wu,Kun Kuang

Long-tailed distribution of semantic categories, which has been often ignored in conventional methods, causes unsatisfactory performance in semantic segmentation on tail categories. In this paper, we focus on the problem of long-tailed semantic segmentation. Although some long-tailed recognition methods (e.g., re-sampling/re-weighting) have been proposed in other problems, they are likely to compromise crucial contextual information in semantic segmentation. Therefore, these methods are hardly adaptable to the problem of long-tailed semantic segmentation. To address this problem, we propose a novel method, named MEDOE, by ensembling and grouping contextual information. Specifically, our MEDOE is a two-sage fram ework comprising a multi-expert decoder (MED) and a multi-expert output ensemble (MOE). The MED includes several ``experts", each of which takes as input the dataset masked according to the specific categories based on frequency distribution and generates contextual information self-adaptively for classification. The MOE then ensembles the experts' outputs with learnable decision weights. As a model-agnostic framework, MEDOE can be flexibly and efficiently coupled with various popular deep neural networks (e.g., Deeplabv3+, OCRNet, and PSPNet) to improve the performance in long-tailed semantic segmentation. Experimental results show that the proposed framework outperforms the current methods on both Cityscapes and ADE20K datasets by up to 2\% in mIoU and 6\% in mAcc.

****

Concept-level Debugging of Part-Prototype Networks
Andrea Bontempelli,Stefano Teso,Katya Tentori,Fausto Giunchiglia,Andrea Passerini

Part-prototype Networks (ProtoPNets) are concept-based classifiers designed to achieve the same performance as black-box models without compromising transparency. ProtoPNets compute predictions based on similarity to class-specific part-prototypes learned to recognize parts of training examples, making it easy to faithfully determine what examples are responsible for any target prediction and why. However, like other models, they are prone to picking up confounders and shortcuts from the data, thus suffering from compromised prediction accuracy and limited generalization. We propose ProtoPDebug, an effective concept-level debugger for ProtoPNets in which a human supervisor, guided by the model’s explanations, supplies feedback in the form of what part-prototypes must be forgotten or kept, and the model is fine-tuned to align with this supervision. Our experimental evaluation shows that ProtoPDebug outperforms state-of-the-art debuggers for a fraction of the annotation cost. An online experiment with laypeople confirms the simplicity of the feedback requested to the users and the effectiveness of the collected feedback for learning confounder-free part-prototypes. ProtoPDebug is a promising tool for trustworthy interactive learning in critical applications, as suggested by a preliminary evaluation on a medical decision making task.

****

Geometrically regularized autoencoders for non-Euclidean data
Cheongjae Jang,Yonghyeon Lee,Yung-Kyun Noh,Frank C. Park

Regularization is almost {\it de rigueur} when designing autoencoders that are sparse and robust to noise. Given the recent surge of interest in machine learning problems involving non-Euclidean data, in this paper we address the regularization of autoencoders on curved spaces. We show that by ignoring the underlying geometry of the data and applying standard vector space regularization techniques, autoencoder performance can be severely degraded, or worse, training can fail to converge. Assuming that both the data space and latent space can be modeled as Riemannian manifolds, we show how to construct regularization terms in a coordinate-invariant way, and develop geometric generalizations of the denoising autoencoder and reconstruction contractive autoencoder such that the essential properties that enable the estimation of the derivative of the log-probability density are preserved. Drawing upon various non-Euclidean data sets, we show that our geometric autoencoder regularization techniques can have important performance advantages over vector-spaced methods while avoiding other breakdowns that can result from failing to account for the underlying geometry.

****

TransFool: An Adversarial Attack against Neural Machine Translation Models
Sahar Sadrizadeh,Pascal Frossard,Ljiljana Dolamic

Deep neural networks have been shown to be vulnerable to small perturbations of their inputs known as adversarial attacks. In this paper, we consider the particular task of Neural Machine Translation (NMT), where security is often critical. We investigate the vulnerability of NMT models to adversarial attacks and propose a new attack algorithm called TransFool. It builds on a multi-term optimization problem and a gradient projection step to compute adversarial examples that fool NMT models. By integrating the embedding representation of a language model in the proposed attack, we generate fluent adversarial examples in the source la

nguage that maintain a high level of semantic similarity with the clean samples and render the attack largely undetectable. Experimental results demonstrate that, for multiple translation tasks and different NMT architectures, our white-box attack can severely degrade the translation quality for more than 60% of the sentences while the semantic similarity between the original sentence and the adversarial example stays very high. Moreover, we show that the proposed attack is transferable to unknown target models and can fool those quite easily. Finally, our method leads to improvement in terms of success rate, semantic similarity, and fluency compared to the existing attack strategies both in white-box and black-box settings. Hence, TransFool permits to better characterize the vulnerability of NMT systems and outlines the necessity to design strong defense mechanisms and more robust NMT systems for real-life applications.
****

Protein Sequence Design in a Latent Space via Model-based Reinforcement Learning
Minji Lee,Luiz Felipe Vecchietti,Hyunkyu Jung,Hyunjoo Ro,Meeyoung Cha,Ho Min Kim
Proteins are complex molecules responsible for different functions in the human body. Enhancing the functionality of a protein and/or cellular fitness can significantly impact various industries. However, their optimization remains challenging, and sequences generated by data-driven methods often fail in wet lab experiments. This study investigates the limitations of existing model-based sequence design methods and presents a novel optimization framework that can efficiently traverse the latent representation space instead of the protein sequence space. Our framework generates proteins with higher functionality and cellular fitness by modeling the sequence design task as a Markov decision process and applying model-based reinforcement learning. We discuss the results in a comprehensive evaluation of two distinct proteins, GPF and His3, along with the predicted structure of optimized sequences using deep learning-based structure prediction.
****

The GANfather: Controllable generation of malicious activity to expose detection weaknesses and improve defence systems.
Ricardo Ribeiro Pereira,Jacopo Bono,João Tiago Ascensão,David Oliveira Aparicio,Pedro Manuel Pinto Ribeiro,Pedro Bizarro
Criminal activities are typically adversarial in nature, where an attacker and a defence system are constantly adapting to each other's behaviour. If the defence systems are helped by automated detection methods, then those methods need to be updated frequently. In practice, this means that the defence systems are always one step behind the attackers. For example, in anti-money laundering systems, new labels representing suspicious activity are frequently delayed by weeks or months and some money laundering activity may never be found, leading to detection systems that are inaccurate and resulting in an estimated undetected €0.7-3 trillion being laundered annually.

To tackle the problem of missing or delayed labels in adversarial settings, we propose The GANfather, an adversarial and label-free method to both (1) generate a variety of meaningful attacks, as guided by a custom, user-defined objective function; and (2) train a defence system to detect such attacks. Optionally, we can ensure that the generated attacks escape an existing detection system, revealing current weaknesses which the new defence system actively corrects. Our method is inspired by generative adversarial networks (GANs), but unlike GANs we nudge our generator to produce out-of-distribution data using a loss function that characterises criminal activity. Importantly, our method does not require any labelled examples.

We test our framework in two real-world use-cases, namely injection attacks in recommendation systems and anti-money laundering. In the former, we show how an injection attack with a limited number of generated fake profiles is sufficient to successfully recommend an item to a large number of users. These generated injection attacks are more effective in recommending the target item than naive ‘bombing’ strategies and harder to detect. In the latter, the generated attacks are able to simulate money laundering and move cumulative amounts close to 250 thou sand dollars through a network of accounts without being detected by existing systems. We also show how we can train a new defence system that captures all these synthetic attacks, potentially saving millions of dollars in detected criminal activity. Our method is generic and applicable in a variety of adversarial domains, exposing current liabilities with the generated data and strengthening the defence systems against current and future malicious attacks.

****

Proximal Validation Protocol

MingFeng Ou,Yiming Zhang,Sai Wu,Gang Chen,Junbo Zhao

Modern machine learning algorithms are generally built upon a train/validation/test split protocol. In particular, with the absence of accessible testing set in real-world ML development, how to split out a validation set becomes crucial for reliable model evaluation, selection and etc. Concretely, under a randomized splitting setup, the split ratio of the validation set generally acts as a vital meta-parameter; that is, with more data picked and used for validation, it would cost model performance due to the less training data, and vice versa. Unfortunately, this implies a vexing trade-off between performance enhancement against trustful model evaluation. However, to date, the research conducted on this line remains very few. We reason this could be due to a workflow gap between the academic and ML production which we may attribute to a form of technical debt of ML. In this article, we propose a novel scheme --- dubbed Proximal Validation Protocol (PVP) --- which is targeted to resolve this problem of validation set construction. Core to PVP is to assemble a \emph{proximal set} as a substitution for the traditional validation set while avoiding the valuable data wasted by the training procedure. The construction of the proximal validation set is established with dense data augmentation followed by a novel distributional-consistent sampling algorithm. With extensive empirical findings, we prove that PVP works (much) better than all the other existing validation protocols on three data modalities (images, text, and tabular data), demonstrating its feasibility towards ML production.

****

A Message Passing Perspective on Learning Dynamics of Contrastive Learning

Yifei Wang,Qi Zhang,Tianqi Du,Jiansheng Yang,Zhouchen Lin,Yisen Wang

In recent years, contrastive learning achieves impressive results on self-supervised visual representation learning, but there still lacks a rigorous understanding of its learning dynamics. In this paper, we show that if we cast a contrastive objective equivalently into the feature space, then its learning dynamics admits an interpretable form. Specifically, we show that its gradient descent corresponds to a specific message passing scheme on the corresponding augmentation graph. Based on this perspective, we theoretically characterize how contrastive learning gradually learns discriminative features with the alignment update and the uniformity update. Meanwhile, this perspective also establishes an intriguing connection between contrastive learning and Message Passing Graph Neural Networks (MP-GNNs). This connection not only provides a unified understanding of many techniques independently developed in each community, but also enables us to borrow techniques from MP-GNNs to design new contrastive learning variants, such as graph attention, graph rewiring, jumpy knowledge techniques, etc. We believe that our message passing perspective not only provides a new theoretical understanding of contrastive learning dynamics, but also bridges the two seemingly independent areas together, which could inspire more interleaving studies to benefit from each other. The code is available at https://github.com/PKU-ML/Message-Passing-Contrastive-Learning.

****

Help Me Explore: Combining Autotelic and Social Learning via Active Goal Queries

Ahmed Akakzia,Hugo Caselles-Dupré,Olivier Serris,Mohamed CHETOUANI,Olivier Sigaud,Cédric Colas

Most approaches to open-ended skill learning train a single agent in a purely sensorimotor environment. But because no human child learns everything on their own, we argue that sociality will be a key component of open-ended learning systems. This paper enables learning agents to blend individual and socially-guided sk ill learning through a new interaction protocol named Help Me Explore (HME). In social episodes triggered at the agent's demand, a social partner suggests a goal at the frontier of the agent's capabilities and, when the goal is reached, follows up with a new adjacent goal just beyond. In individual episodes, the agent practices skills autonomously by pursuing goals it already discovered through either its own experience or social suggestions. The idea of augmenting an individual goal exploration with social goal suggestions is simple, general and powerful. We demonstrate its efficiency on two notoriously hard exploration benchmarks: continuous mazes and a 5-block robotic manipulation task. With minimal social interventions, an HME-agent outperforms the purely social agent deprived of its autonomy, and the purely individual agent which fails to solve hard exploration problems.

****

AUTOMATIC CURRICULUM FOR UNSUPERVISED REIN- FORCEMENT LEARNING

Yucheng Yang,Tianyi Zhou,Tianhong Dai,Meng Fang,Mykola Pechenizkiy

Recent unsupervised reinforcement learning (URL) can learn meaningful skills without task rewards by carefully designed training objectives. However, most existing works lack quantitative evaluation metrics for URL but mainly rely on visualizations of trajectories to compare the performance. Moreover, each URL method only focuses on a single training objective, which can hinder further learning progress and the development of new skills. To bridge these gaps, we first propose multiple evaluation metrics for URL that can cover different preferred properties. We show that balancing these metrics leads to what a “good” trajectory visualization embodies. Next, we use these metrics to develop an automatic curriculum that can change the URL objective across different learning stages in order to improve and balance all metrics. Specifically, we apply a non-stationary multi-armed bandit algorithm to select an existing URL objective for each episode according to the metrics evaluated in previous episodes. Extensive experiments indifferent environments demonstrate the advantages of our method on achieving promising and balanced performance over all URL metrics.

****

Filtered Semi-Markov CRF

Urchade Zaratiana,Nadi Tomeh,Niama Elkhbir,Pierre Holat,Thierry Charnois

Semi-Markov CRF \citep{semicrf} has been proposed as an alternative to the traditional Linear Chain CRF\citep{crf} for text segmentation tasks such as Named Entity Recognition. In contrast to CRF, which treats text segmentation as token-level prediction, Semi-CRF considers spans as the task's basic unit, which makes it more expressive. However, Semi-CRF has two major drawbacks: (1) it has quadratic complexity over sequence length as it operates on every span of the input sequence, and (2) empirically, it performs worse than classical CRF for sequence labeling tasks such as NER. In our work, we propose Filtered Semi-Markov CRF, a Semi-CRF variant that addresses the aforementioned issues. Our model extends Semi-CRF by incorporating a filtering step for eliminating irrelevant segments, which helps in reducing the complexity and allows to dramatically reduce the search space. On a variety of NER benchmarks, we find that our approach outperforms both CRF and Semi-CRF models while being significantly faster. We will make our code available to the public.

****

Zeroth-Order Optimization with Trajectory-Informed Derivative Estimation

Yao Shu,Zhongxiang Dai,Weicong Sng,Arun Verma,Patrick Jaillet,Bryan Kian Hsiang Low

Zeroth-order (ZO) optimization, in which the derivative is unavailable, has recently succeeded in many important machine learning applications. Existing algorithms rely on finite difference (FD) methods for derivative estimation and gradient descent (GD)-based approaches for optimization. However, these algorithms suffer from query inefficiency because many additional function queries are required for derivative estimation in their every GD update, which typically hinders their deployment in real-world applications where every function query is expensive. To this end, we propose a trajectory-informed derivative estimation method which only employs the optimization trajectory (i.e., the history of function queri es during optimization) and hence can eliminate the need for additional function queries to estimate a derivative. Moreover, based on our derivative estimation, we propose the technique of dynamic virtual updates, which allows us to reliably perform multiple steps of GD updates without reapplying derivative estimation. Based on these two contributions, we introduce the zeroth-order optimization with trajectory-informed derivative estimation (ZoRD) algorithm for query-efficient ZO optimization. We theoretically demonstrate that our trajectory-informed derivative estimation and our ZoRD algorithm improve over existing approaches, which is then supported by our real-world experiments such as black-box adversarial attack, non-differentiable metric optimization, and derivative-free reinforcement learning.

****

Distance VS. Coordinate: Distance Based Embedding Improves Model Generalization for Routing Problems

Hongsen Liao,Ruiyuan Wu,Yuyang Han,Yuncong Hu,Ke Xing,Jinghua Hao,Renqing He

Routing problems, such as traveling salesman problem (TSP) and vehicle routing problem, are among the most classic research topics in combinatorial optimization and operations research (OR). In recent years, with the rapid development of online service platforms, there has been renewed interest in applying this study to facilitate emerging industrial applications, such as food delivery and logistics services. While OR methods remain the mainstream technique, increasing efforts have been put into exploiting deep learning (DL) models for tackling routing problems. The existing ML methods often consider the embedding of the route point coordinate as a key model input and are capable of delivering competing performance in synthetic or simplified settings. However, it is empirically noted that this line of work appears to lack robustness and generalization ability that are crucial for real-world applications. In this paper, we demonstrate that the coordinate can unexpectedly lead to these problems. There are two factors that make coordinate rather `poisonous' for DL models: i) the definition of distance between route points is far more complex than what coordinate can depict; ii) the coordinate can hardly be sufficiently `traversed' by the training data. To circumvent these limitations, we propose to abandon the coordinate and instead use the relative distance for route point embedding. We show in both synthetic TSP and real-world food pickup and delivery route prediction problem that our design can significantly improve model's generalization ability, and deliver competitive or better performance with existing models.

****

Towards biologically plausible Dreaming and Planning

Cristiano Capone,Pier Stanislao Paolucci

Humans and animals can learn new skills after practicing for a few hours, while current reinforcement learning algorithms require a large amount of data to achieve good performances.

Recent model-based approaches show promising results by reducing the number of necessary interactions with the environment to learn a desirable policy. However, these methods require biological implausible ingredients, such as the detailed storage of older experiences, and long periods of offline learning. The optimal way to learn and exploit word-models is still an open question.

Taking inspiration from biology, we suggest that dreaming might be an efficient expedient to use an inner model. We propose a two-module (agent and model) neural network in which "dreaming" (living new experiences in a model-based simulated environment) significantly boosts learning. We also explore "planning", an online alternative to dreaming, that shows comparable performances. Importantly, our model does not require the detailed storage of experiences, and learns online the world-model. This is a key ingredient for biological plausibility and implementability (e.g., in neuromorphic hardware).

****

Extracting Meaningful Attention on Source Code: An Empirical Study of Developer and Neural Model Code Exploration

Matteo Paltenghi,Rahul Pandita,Austin Henley,Albert Ziegler

The high effectiveness of neural models of code, such as OpenAI Codex and AlphaC ode, suggests coding capabilities of models that are at least comparable to those of humans. However, previous work has only used these models for their raw completion, ignoring how the model reasoning, in the form of attention weights, can be used for other downstream tasks. Disregarding the attention weights means discarding a considerable portion of what those models compute when queried. To profit more from the knowledge embedded in these large pre-trained models, this work compares multiple approaches to post-process these valuable attention weights for supporting code exploration. Specifically, we compare to which extent the transformed attention signal of CodeGen, a large and publicly available pre-trained neural model, agrees with how developers look at and explore code when each answering the same sense-making questions about code. At the core of our experimental evaluation, we collect, manually annotate, and open-source a novel eye-tracking dataset comprising 25 developers answering sense-making questions on code over 92 sessions. We empirically evaluate five attention-agnostic heuristics and ten attention-based post processing approaches of the attention signal against our ground truth of developers exploring code, including the novel concept of follow-up attention which exhibits the highest agreement. Beyond the dataset contribution and the empirical study, we also introduce a novel practical application of the attention signal of pre-trained models with completely analytical solutions, going beyond how neural models’ attention mechanisms have traditionally been used.

****

Neuroevolution is a Competitive Alternative to Reinforcement Learning for Skill Discovery

Felix Chalumeau,Raphael Boige,Bryan Lim,Valentin Macé,Maxime Allard,Arthur Flajolet,Antoine Cully,Thomas PIERROT

Deep Reinforcement Learning (RL) has emerged as a powerful paradigm for training neural policies to solve complex control tasks. However, these policies tend to be overfit to the exact specifications of the task and environment they were trained on, and thus do not perform well when conditions deviate slightly or when composed hierarchically to solve even more complex tasks. Recent work has shown that training a mixture of policies, as opposed to a single one, that are driven to explore different regions of the state-action space can address this shortcoming by generating a diverse set of behaviors, referred to as skills, that can be collectively used to great effect in adaptation tasks or for hierarchical planning. This is typically realized by including a diversity term - often derived from information theory - in the objective function optimized by RL. However these approaches often require careful hyperparameter tuning to be effective. In this work, we demonstrate that less widely-used neuroevolution methods, specifically Quality Diversity (QD), are a competitive alternative to information-theory-augmented RL for skill discovery. Through an extensive empirical evaluation comparing eight state-of-the-art algorithms (four flagship algorithms from each line of work) on the basis of (i) metrics directly evaluating the skills' diversity, (ii) the skills' performance on adaptation tasks, and (iii) the skills' performance when used as primitives for hierarchical planning; QD methods are found to provide equal, and sometimes improved, performance whilst being less sensitive to hyperparameters and more scalable. As no single method is found to provide near-optimal performance across all environments, there is a rich scope for further research which we support by proposing future directions and providing optimized open-source implementations.

****

Scrunch: Preventing sensitive property inference through privacy-preserving representation learning

Vittorio Prodomo,Roberto González-Sánchez,Marco Gramaglia

Many tasks that are commonly performed by devices attached to the Internet are currently being offloaded to the cloud, using the Machine Learning as a Service (MLaaS) paradigm. While this paradigm is motivated by the reduced capacity of mobile terminals, it also hinders privacy associated with the data exchanged over the network. Thus, the data exchanged among parties shall be conveniently anonymi zed to prevent possible confidentiality and privacy issues. While many privacy-enhancing algorithms have been proposed in the past, they are usually relying on very complex models that make difficult their applicability to real-world systems or envision too friendly attacker models. In this paper, we propose a deep learning system that creates anonymized representations for the data, while keeping the accuracy for the targeted MLaaS task high, assuming that the attacker can re-train an adversarial model. Our results show that the proposed algorithm i) is effective yet it uses a lighter approach than state-of-the-art ii) considers less friendly attacker models, and iii) outperforms the benchmark under different privacy metrics.

****

Uniform-in-time propagation of chaos for the mean field gradient Langevin dynamics

Taiji Suzuki,Atsushi Nitanda,Denny Wu

The mean-field Langevin dynamics is characterized by a stochastic differential equation that arises from (noisy) gradient descent on an infinite-width two-layer neural network, which can be viewed as an interacting particle system. In this work, we establish a quantitative weak propagation of chaos result for the system, with a finite-particle discretization error of $\mathcal{O}(1/N)$ \textit{uniformly over time}, where N is the width of the neural network. This allows us to directly transfer the optimization guarantee for infinite-width networks to practical finite-width models without excessive overparameterization. On the technical side, our analysis differs from most existing studies on similar mean field dynamics in that we do not require the interaction between particles to be sufficiently weak to obtain a uniform propagation of chaos, because such assumptions may not be satisfied in neural network optimization. Instead, we make use of a logarithmic Sobolev-type condition which can be verified in appropriate regularized risk minimization settings.

****

GM-VAE: Representation Learning with VAE on Gaussian Manifold

Seunghyuk Cho,Juyong Lee,Dongwoo Kim

We propose a Gaussian manifold variational auto-encoder (GM-VAE) whose latent space consists of a set of diagonal Gaussian distributions. It is known that the set of the diagonal Gaussian distributions with the Fisher information metric forms a product hyperbolic space, which we call a Gaussian manifold. To learn the VAE endowed with the Gaussian manifold, we first propose a pseudo Gaussian manifold normal distribution based on the Kullback-Leibler divergence, a local approximation of the squared Fisher-Rao distance, to define a density over the latent space. With the newly proposed distribution, we introduce geometric transformations at the last and the first of the encoder and the decoder of VAE, respectively to help the transition between the Euclidean and Gaussian manifolds. Through the empirical experiments, we show competitive generalization performance of GM-VAE against other variants of hyperbolic- and Euclidean-VAEs. Our model achieves strong numerical stability, which is a common limitation reported with previous hyperbolic-VAEs.

****

Improving Adversarial Robustness by Putting More Regularizations on Less Robust Samples

Dongyoon Yang,Insung Kong,Yongdai Kim

Adversarial training, which is to enhance robustness against adversarial attacks, has received much attention because it is easy to generate human-imperceptible perturbations of data to deceive a given deep neural network. In this paper, we propose a new adversarial training algorithm that is theoretically well motivated and empirically superior to other existing algorithms. A novel feature of the proposed algorithm is to apply more regularization to data vulnerable to adversarial attacks than other existing regularization algorithms do. Theoretically, we show that our algorithm can be understood as an algorithm of minimizing a newly derived upper bound of the robust risk. Numerical experiments illustrate that our proposed algorithm improves the generalization (accuracy on examples) and ro bustness (accuracy on adversarial attacks) simultaneously to achieve the state-of-the-art performance.

****

Causal Explanations of Structural Causal Models

Matej Ze■evi■,Devendra Singh Dhami,Constantin A. Rothkopf,Kristian Kersting

In explanatory interactive learning (XIL) the user queries the learner, then the learner explains its answer to the user and finally the loop repeats. XIL is attractive for two reasons, (1) the learner becomes better and (2) the user's trust increases. For both reasons to hold, the learner's explanations must be useful to the user and the user must be allowed to ask useful questions. Ideally, both questions and explanations should be grounded in a causal model since they avoid spurious fallacies. Ultimately, we seem to seek a causal variant of XIL. The question part on the user's end we believe to be solved since the user's mental model can provide the causal model. But how would the learner provide causal explanations? In this work we show that existing explanation methods are not guaranteed to be causal even when provided with a Structural Causal Model (SCM). Specifically, we use the popular, proclaimed causal explanation method CXPlain to illustrate how the generated explanations leave open the question of truly causal explanations. Thus as a step towards causal XIL, we propose a solution to the lack of causal explanations. We solve this problem by deriving from first principles an explanation method that makes full use of a given SCM, which we refer to as SC$\textbf{E}$ ($\textbf{E}$ standing for explanation). Since SCEs make use of structural information, any causal graph learner can now provide human-readable explanations. We conduct several experiments including a user study with 22 participants to investigate the virtue of SCE as causal explanations of SCMs.

****

Asynchronous Distributed Bilevel Optimization

Yang Jiao,Kai Yang,Tiancheng Wu,Dongjin Song,Chengtao Jian

Bilevel optimization plays an essential role in many machine learning tasks, ranging from hyperparameter optimization to meta-learning. Existing studies on bilevel optimization, however, focus on either centralized or synchronous distributed setting. The centralized bilevel optimization approaches require collecting massive amount of data to a single server, which inevitably incur significant communication expenses and may give rise to data privacy risks. Synchronous distributed bilevel optimization algorithms, on the other hand, often face the straggler problem and will immediately stop working if a few workers fail to respond. As a remedy, we propose Asynchronous Distributed Bilevel Optimization (ADBO) algorithm. The proposed ADBO can tackle bilevel optimization problems with both nonconvex upper-level and lower-level objective functions, and its convergence is theoretically guaranteed. Furthermore, it is revealed through theoretic analysis that the iteration complexity of ADBO to obtain the ϵ-stationary point is upper bounded by $\mathcal{O}(\frac{1}{{{\epsilon ^2}}})$. Thorough empirical studies on public datasets have been conducted to elucidate the effectiveness and efficiency of the proposed ADBO.

****

Multi-Agent Reinforcement Learning with Shared Resources for Inventory Management

Yuandong Ding,Mingxiao Feng,Guozi Liu,Wei Jiang,Chuheng Zhang,Li Zhao,Lei Song,Houqiang Li,Yan Jin,Jiang Bian

In this paper, we consider the inventory management (IM) problem where we need to make replenishment decisions for a large number of stock keeping units (SKUs) to balance their supply and demand. In our setting, the constraint on the shared resources (such as the inventory capacity) couples the otherwise independent control for each SKU. We formulate the problem with this structure as Shared-Resource Stochastic Game (SRSG) and propose an efficient algorithm called Context-aware Decentralized PPO (CD-PPO). Through extensive experiments, we demonstrate that CD-PPO can accelerate the learning procedure compared with standard MARL algorithms.

****

Confidence-Based Feature Imputation for Graphs with Partially Known Features

Daeho Um,Jiwoong Park,Seulki Park,Jin young Choi

This paper investigates a missing feature imputation problem for graph learning tasks. Several methods have previously addressed learning tasks on graphs with missing features. However, in cases of high rates of missing features, they were unable to avoid significant performance degradation. To overcome this limitation, we introduce a novel concept of channel-wise confidence in a node feature, which is assigned to each imputed channel feature of a node for reflecting the certainty of the imputation. We then design pseudo-confidence using the channel-wise shortest path distance between a missing-feature node and its nearest known-feature node to replace unavailable true confidence in an actual learning process. Based on the pseudo-confidence, we propose a novel feature imputation scheme that performs channel-wise inter-node diffusion and node-wise inter-channel propagation. The scheme can endure even at an exceedingly high missing rate (e.g., 99.5\%) and it achieves state-of-the-art accuracy for both semi-supervised node classification and link prediction on various datasets containing a high rate of missing features. Codes are available at https://github.com/daehoum1/pcfi.

****

Explicitly Maintaining Diverse Playing Styles in Self-Play

Yuan Liu,Ruimin Shen,Miqing Li,Yingfeng Chen,Juan Zou,Changjie Fan

Self-play has proven to be an effective training schema to obtain a high-level agent in complex games through iteratively playing against an opponent from its historical versions. However, its training process may prevent it from generating a well-generalised policy since the trained agent rarely encounters diversely-behaving opponents along its own historical path. In this paper, we aim to improve the generalisation of the policy by maintaining a population of agents with diverse playing styles and high skill levels throughout the training process. Specifically, we propose a bi-objective optimisation model to simultaneously optimise the agents' skill level and playing style. A feature of this model is that we do not regard the skill level and playing style as two objectives to maximise directly since they are not equally important (i.e., agents with diverse playing styles but low skill levels are meaningless). Instead, we create a meta bi-objective model to enable high-level agents with diverse playing styles more likely to be incomparable (i.e. Pareto non-dominated), thereby playing against each other through the training process. We then present an evolutionary algorithm working with the proposed model. Experiments in a classic table tennis game Pong and a commercial role-playing game Justice Online show that our algorithm can learn a well generalised policy and at the same time is able to provide a set of high-level policies with various playing styles.

****

Toward Learning Geometric Eigen-Lengths Crucial for Robotic Fitting Tasks

Yijia Weng,Kaichun Mo,Ruoxi Shi,Yanchao Yang,Leonidas Guibas

Some extremely low-dimensional yet crucial geometric eigen-lengths often determine whether an object can be fitted in the environment or not. For example, the {\em height} of an object is important to measure to check if it can fit between the shelves of a cabinet, while the {\em width} of a couch is crucial when trying to move it through a doorway. Humans have materialized such crucial geometric eigen-lengths in common sense since they are very useful in serving as succinct yet effective, highly interpretable, and universal object representations. However, it remains obscure and underexplored if learning systems can be equipped with similar capabilities of automatically discovering such key geometric quantities in doing robotic fitting tasks. In this work, we therefore for the first time formulate and propose a novel learning problem on this question and set up a benchmark suite including the tasks, the data, and the evaluation metrics for studying the problem. We explore potential solutions and demonstrate the feasibility of learning such eigen-lengths from simply observing successful and failed fitting trials. We also attempt geometric grounding for more accurate eigen-length measurement and study the reusability of the learned geometric eigen-lengths across multiple tasks. Our work marks the first exploratory step toward learning crucial geometric eigen-lengths and we hope it can inspire future research in tackling this important yet underexplored problem.

****

LiftedCL: Lifting Contrastive Learning for Human-Centric Perception

Ziwei Chen,Qiang Li,Xiaofeng Wang,Wankou Yang

Human-centric perception targets for understanding human body pose, shape and segmentation. Pre-training the model on large-scale datasets and fine-tuning it on specific tasks has become a well-established paradigm in human-centric perception. Recently, self-supervised learning methods have re-investigated contrastive learning to achieve superior performance on various downstream tasks. When handling human-centric perception, there still remains untapped potential since 3D human structure information is neglected during the task-agnostic pre-training. In this paper, we propose the Lifting Contrastive Learning (LiftedCL) to obtain 3D-aware human-centric representations which absorb 3D human structure information. In particular, to induce the learning process, a set of 3D skeletons is randomly sampled by resorting to 3D human kinematic prior. With this set of generic 3D samples, 3D human structure information can be learned into 3D-aware representations through adversarial learning. Empirical results demonstrate that LiftedCL outperforms state-of-the-art self-supervised methods on four human-centric downstream tasks, including 2D and 3D human pose estimation (0.4% mAP and 1.8 mm MPJPE improvement on COCO 2D pose estimation and Human3.6M 3D pose estimation), human shape recovery and human parsing.

****

Individual Privacy Accounting with Gaussian Differential Privacy

Antti Koskela,Marlon Tobaben,Antti Honkela

Individual privacy accounting enables bounding differential privacy (DP) loss individually for each participant involved in the analysis. This can be informative as often the individual privacy losses are considerably smaller than those indicated by the DP bounds that are based on considering worst-case bounds at each data access. In order to account for the individual losses in a principled manner, we need a privacy accountant for adaptive compositions of mechanisms, where the loss incurred at a given data access is allowed to be smaller than the worst-case loss. This kind of analysis has been carried out for the Rényi differential privacy by Feldman and Zrnic (2021), however not yet for the so-called optimal privacy accountants. We make first steps in this direction by providing a careful analysis using the Gaussian differential privacy which gives optimal bounds for the Gaussian mechanism, one of the most versatile DP mechanisms. This approach is based on determining a certain supermartingale for the hockey-stick divergence and on extending the Rényi divergence-based fully adaptive composition results by Feldman and Zrnic (2021). We also consider measuring the individual (ε,δ)-privacy losses using the so-called privacy loss distributions. Using the Blackwell theorem, we can then use the results of Feldman and Zrnic (2021) to construct an approximative individual (ε,δ)-accountant. We also show how to speed up the FFT-based individual DP accounting using the Plancherel theorem.

****

Evolving Populations of Diverse RL Agents with MAP-Elites

Thomas PIERROT,Arthur Flajolet

Quality Diversity (QD) has emerged as a powerful alternative optimization paradigm that aims at generating large and diverse collections of solutions, notably with its flagship algorithm MAP-ELITES (ME) which evolves solutions through mutations and crossovers. While very effective for some unstructured problems, early ME implementations relied exclusively on random search to evolve the population of solutions, rendering them notoriously sample-inefficient for high-dimensional problems, such as when evolving neural networks. Follow-up works considered exploiting gradient information to guide the search in order to address these shortcomings through techniques borrowed from either Black-Box Optimization (BBO) or Reinforcement Learning (RL). While mixing RL techniques with ME unlocked state-of-the-art performance for robotics control problems that require a good amount of exploration, it also plagued these ME variants with limitations common among R

L algorithms that ME was free of, such as hyperparameter sensitivity, high stochasticity as well as training instability, including when the population size increases as some components are shared across the population in recent approaches. Furthermore, existing approaches mixing ME with RL tend to be tied to a specific RL algorithm, which effectively prevents their use on problems where the corresponding RL algorithm fails. To address these shortcomings, we introduce a flexible framework that allows the use of any RL algorithm and alleviates the aforementioned limitations by evolving populations of agents (whose definition include hyperparameters and all learnable parameters) instead of just policies. We demonstrate the benefits brought about by our framework through extensive numerical experiments on a number of robotics control problems, some of which with deceptive rewards, taken from the QD-RL literature. We open source an efficient JAX-based implementation of our algorithm in the QDax library.

****

Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data
Spencer Frei,Gal Vardi,Peter Bartlett,Nathan Srebro,Wei Hu

The implicit biases of gradient-based optimization algorithms are conjectured to be a major factor in the success of modern deep learning. In this work, we investigate the implicit bias of gradient flow and gradient descent in two-layer fully-connected neural networks with leaky ReLU activations when the training data are nearly-orthogonal, a common property of high-dimensional data. For gradient flow, we leverage recent work on the implicit bias for homogeneous neural networks to show that asymptotically, gradient flow produces a neural network with rank at most two. Moreover, this network is an ℓ_2-max-margin solution (in parameter space), and has a linear decision boundary that corresponds to an approximate-max-margin linear predictor. For gradient descent, provided the random initialization variance is small enough, we show that a single step of gradient descent suffices to drastically reduce the rank of the network, and that the rank remains small throughout training. We provide experiments which suggest that a small initialization scale is important for finding low-rank neural networks with gradient descent.

****

Learning Test Time Augmentation with Cascade Loss Prediction
Siyang Pan,Jiaqian Yu,Dongwook Lee,Qiang Wang,ChangBeom Park,ByungIn Yoo

Data augmentation has been a successful common practice for improving the performance of deep neural network during training stage. In recent years, studies on test time augmentation (TTA) have also been promising due to its effectiveness on improving the robustness against out-of-distribution data at inference. Instead of simply adopting pre-defined handcrafted geometric operations such as croping and flipping, recent TTA methods learn predictive transformations which are supposed to provide the best performance gain on each test sample. However, the desired iteration number of transformation is proportional to the inference time of the predictor, and the gain by ensembling multiple augmented inputs still requires additional forward pass of the target model. In this paper, we propose a cascade method for test time augmentation prediction. It only requires a single forward pass of the transformation predictor, while can output multiple desirable transformations iteratively. These transformations will then be adopted sequentially on the test sample at once before the target model inference. The experimental results show that our method provides a better trade-off between computational cost and overall performance at test time, and shows significant improvement compared to existing methods.

****

Adaptive Computation with Elastic Input Sequence
Fuzhao Xue,Valerii Likhosherstov,Anurag Arnab,Neil Houlsby,Yi Tay,Mostafa Dehghani,Yang You

When solving a problem, human beings have the adaptive ability in terms of the type of information they use, the procedure they take, and the amount of time they spend approaching and solving the problem. However, most standard neural networks have the same function type and fixed computation budget on different samples regardless of their nature and difficulty. Adaptivity is a powerful paradigm a s it not only imbues practitioners with flexibility pertaining to the downstream usage of these models but can also serve as a powerful inductive bias for solving certain challenging classes of problems. In this work, we propose a new strategy, AdaTape, that enables dynamic computation in neural networks via adaptive tape tokens. AdaTape employs an elastic input sequence by equipping an existing architecture with a dynamic read and write tape. Specifically, we adaptively generate input sequences using tape tokens obtained from a tape bank that can either be trainable or generated from input data. We analyze the challenges and requirements to obtain dynamic sequence content and length, and propose the Adaptive Tape Reader (ATR) algorithm to achieve both objectives. Via extensive experiments on image recognition tasks, we show that AdaTape can achieve better performance while maintaining the computational cost.

****

Optimizing Data-Flow in Binary Neural Networks

Lorenzo Vorabbi,Davide Maltoni,Stefano Santi

Binary Neural Networks (BNNs) can significantly accelerate the inference time of a neural network by replacing its expensive floating-point arithmetic with bit-wise operations. Most existing solutions, however, do not fully optimize data flow through the BNN layers, and intermediate conversions from 1 to 16/32 bits often further hinder efficiency. We propose a novel training scheme that can increase data flow and parallelism in the BNN pipeline; specifically, we introduce a clipping block that decreases the data-width from 32 bits to 8. Furthermore, we reduce the internal accumulator size of a binary layer, usually kept using 32-bit to prevent data overflow without losing accuracy. Additionally, we provide an optimization of the Batch Normalization layer that both reduces latency and simplifies deployment. Finally, we present an optimized implementation of the Binary Direct Convolution for ARM instruction sets. Our experiments show a consistent improvement of the inference speed (up to 1.77 and $1.9 \times$ compared to two state-of-the-art BNNs frameworks) with no drop in accuracy for at least one full-precision model.

****

Gray-Box Gaussian Processes for Automated Reinforcement Learning

Gresa Shala,André Biedenkapp,Frank Hutter,Josif Grabocka

Despite having achieved spectacular milestones in an array of important real-world applications, most Reinforcement Learning (RL) methods are very brittle concerning their hyperparameters. Notwithstanding the crucial importance of setting the hyperparameters in training state-of-the-art agents, the task of hyperparameter optimization (HPO) in RL is understudied. In this paper, we propose a novel gray-box Bayesian Optimization technique for HPO in RL, that enriches Gaussian Processes with reward curve estimations based on generalized logistic functions. In a very large-scale experimental protocol, comprising 5 popular RL methods (DDPG, A2C, PPO, SAC, TD3), dozens of environments (Atari, Mujoco), and 7 HPO baselines, we demonstrate that our method significantly outperforms current HPO practices in RL.

****

Protein Sequence and Structure Co-Design with Equivariant Translation

Chence Shi,Chuanrui Wang,Jiarui Lu,Bozitao Zhong,Jian Tang

Proteins are macromolecules that perform essential functions in all living organisms. Designing novel proteins with specific structures and desired functions has been a long-standing challenge in the field of bioengineering. Existing approaches generate both protein sequence and structure using either autoregressive models or diffusion models, both of which suffer from high inference costs. In this paper, we propose a new approach capable of protein sequence and structure co-design, which iteratively translates both protein sequence and structure into the desired state from random initialization, based on context features given a priori. Our model consists of a trigonometry-aware encoder that reasons geometrical constraints and interactions from context features, and a roto-translation equivariant decoder that translates protein sequence and structure interdependently. Notably, all protein amino acids are updated in one shot in each translation step, which significantly accelerates the inference process. Experimental results across multiple tasks show that our model outperforms previous state-of-the-art baselines by a large margin, and is able to design proteins of high fidelity as regards both sequence and structure, with running time orders of magnitude less than sampling-based methods.

****

SGD Through the Lens of Kolmogorov Complexity

Gregory Schwartzman

We initiate a thorough study of the dynamics of stochastic gradient descent (SGD) under minimal assumptions using the tools of entropy compression. Specifically, we characterize a quantity of interest which we refer to as the \emph{accuracy discrepancy}. Roughly speaking, this measures the average discrepancy between the model accuracy on batches and large subsets of the entire dataset. We show that if this quantity is sufficiently large, then SGD finds a model which achieves perfect accuracy on the data in $O(1)$ epochs. On the contrary, if the model cannot perfectly fit the data, this quantity must remain below a \emph{global} threshold, which only depends on the size of the dataset and batch.

We use the above framework to lower bound the amount of randomness required to allow (non stochastic) gradient descent to escape from local minimas using perturbations. We show that even if the model is \emph{extremely overparameterized}, at least a linear (in the size of the dataset) number of random bits are required to guarantee that GD escapes local minimas in polynomial time.

****

Learning in temporally structured environments

Matt Jones,Tyler R. Scott,Mengye Ren,Gamaleldin Fathy Elsayed,Katherine Hermann,David Mayo,Michael Curtis Mozer

Natural environments have temporal structure at multiple timescales. This property is reflected in biological learning and memory but typically not in machine learning systems. We advance a multiscale learning method in which each weight in a neural network is decomposed as a sum of subweights with different learning and decay rates. Thus knowledge becomes distributed across different timescales, enabling rapid adaptation to task changes while avoiding catastrophic interference. First, we prove previous models that learn at multiple timescales, but with complex coupling between timescales, are equivalent to multiscale learning via a reparameterization that eliminates this coupling. The same analysis yields a new characterization of momentum learning, as a fast weight with a negative learning rate. Second, we derive a model of Bayesian inference over $1/f$ noise, a common temporal pattern in many online learning domains that involves long-range (power law) autocorrelations. The generative side of the model expresses $1/f$ noise as a sum of diffusion processes at different timescales, and the inferential side tracks these latent processes using a Kalman filter. We then derive a variational approximation to the Bayesian model and show how it is an extension of the multiscale learner. The result is an optimizer that can be used as a drop-in replacement in an arbitrary neural network architecture. Third, we evaluate the ability of these methods to handle nonstationarity by testing them in online prediction tasks characterized by $1/f$ noise in the latent parameters. We find that the Bayesian model significantly outperforms online stochastic gradient descent and two batch heuristics that rely preferentially or exclusively on more recent data. Moreover, the variational approximation performs nearly as well as the full Bayesian model, and with memory requirements that are linear in the size of the network.

****

Identifying Phase Transition Thresholds of Permuted Linear Regression via Message Passing

Hang Zhang,Ping Li

This paper considers the permuted linear regression, i.e., ${\mathbf{Y}} = {\mathbf{\Pi}}^{\natural}{\mathbf{X}}{\mathbf{B}}^{\natural} + {\mathbf{W}}$, where ${\mathbf{Y}} \in \mathbb{R}^{n\times m}, {\mathbf{\Pi}}^{\natural}\in\mathbb{R}^{n\times n}, {\mathbf{X}} \in \mathbb{R}^{n\times p}, {\mathbf{B}}^{\natural}\in

\mathbb{R}^{p\times m}$, and ${\mathbf{W}}\in \mathbb{R}^{n\times m}$ represent the observations, missing (or incomplete) information about ordering, sensing matrix, signal of interests, and additive sensing noise, respectively. As is shown in the previous work, there exists phase transition phenomena in terms of the \emph{signal-to-noise ratio} ($\mathsf{snr}$), number of permuted rows, etc. While all existing works only concern the convergence rates without specifying the associate constants in front of them, we give a precise identification of the phase transition thresholds via the message passing algorithm. Depending on whether the signal ${\mathbf{B}}^{\natural}$ is known or not, we separately identify the corresponding critical points around the phase transition regimes. Moreover, we provide numerical experiments and show the empirical phase transition points are well aligned with theoretical predictions.

****

RandProx: Primal-Dual Optimization Algorithms with Randomized Proximal Updates

Laurent Condat,Peter Richtárik

Proximal splitting algorithms are well suited to solving large-scale nonsmooth optimization problems, in particular those arising in machine learning. We propose a new primal–dual algorithm, in which the dual update is randomized; equivalently, the proximity operator of one of the function in the problem is replaced by a stochastic oracle. For instance, some randomly chosen dual variables, instead of all, are updated at each iteration. Or, the proximity operator of a function is called with some small probability only. A nonsmooth variance-reduction technique is implemented so that the algorithm finds an exact minimizer of the general problem involving smooth and nonsmooth functions, possibly composed with linear operators. We derive linear convergence results in presence of strong convexity; these results are new even in the deterministic case, when our algorithms reverts to the recently proposed Primal–Dual Davis–Yin algorithm. Some randomized algorithms of the literature are also recovered as particular cases (e.g., Point-SAGA). But our randomization technique is general and encompasses many unbiased mechanisms beyond sampling and probabilistic updates, including compression. Since the convergence speed depends on the slowest among the primal and dual contraction mechanisms, the iteration complexity might remain the same when randomness is used. On the other hand, the computation complexity can be significantly reduced. Overall, randomness helps getting faster algorithms. This has long been known for stochastic-gradient-type algorithms, and our work shows that this fully applies in the more general primal–dual setting as well.

****

Preserving Pre-trained Features Helps Calibrate Fine-tuned Language Models

Guande He,Jianfei Chen,Jun Zhu

Large pre-trained language models (PLMs) have demonstrated strong performance on natural language understanding (NLU) tasks through fine-tuning. However, fine-tuned models still suffer from overconfident predictions, especially in out-of-domain settings. In this paper, we tackle the problem of calibrating fine-tuned language models. We demonstrate that the PLMs are well-calibrated on the masked language modeling task with robust predictive confidence under domain shift, yet the fine-tuned models fail to retain such property due to catastrophic forgetting, which impacts the calibration on the downstream classification task. In light of these observations, we evaluate the calibration of several methods that preserve pre-trained features and show that preserving pre-trained features can improve the calibration of fine-tuned language models. Among these methods, our proposed method that encourages the fine-tuned model to learn generative representations with auxiliary language modeling objective achieves competitive accuracy and the lowest expected calibration error compared to several strong baselines under both in-domain and out-of-domain settings on three downstream NLU tasks.

****

A Hierarchical Bayesian Approach to Federated Learning

Minyoung Kim,Timothy Hospedales

We propose a novel hierarchical Bayesian approach to Federated learning (FL), where our models reasonably describe the generative process of clients' local data via hierarchical Bayesian modeling: constituting random variables of local mode

ls for clients that are governed by a higher-level global variate. Interestingly, the variational inference in our Bayesian model leads to an optimization problem whose block-coordinate descent solution becomes a distributed algorithm that is separable over clients and allows them not to reveal their own private data at all, thus fully compatible with FL. We also highlight that our block-coordinate algorithm has particular forms that subsume the well-known FL algorithms including Fed-Avg and Fed-Prox as special cases. That is, we not only justify the previous Fed-Avg and Fed-Prox algorithms whose learning protocols look intuitive but theoretically less underpinned, but also generalise them even further via principled Bayesian approaches. Beyond introducing novel modeling and derivations, we also offer convergence analysis showing that our block-coordinate FL algorithm converges to an (local) optimum of the objective at the rate of $O(1/\sqrt{t})$, the same rate as regular (centralised) SGD, as well as the generalisation error analysis where we prove that the test error of our model on unseen data is guaranteed to vanish as we increase the training data size, thus asymptotically optimal.
****
Neural Representations in Multi-Task Learning guided by Task-Dependent Contexts
Santiago Galella,Salva Ardid
The ability to switch between tasks effectively in response to external stimuli is a hallmark of cognitive control. Our brain is able to filter and to integrate external information to accomplish goal-directed behavior. Task switching occurs rapidly and efficiently, allowing us to perform multiple tasks with ease. In a similar way, deep learning models can be tailored to exhibit multi-task capabilities and achieve high performance across domains. Still, understanding how neural networks make predictions is crucial in many real-world applications.

In this study, we delve into neural representations learned by multi-tasking architectures. Concretely, we compare individual and parallel networks with task switching networks. Task-switching networks leverage task-dependent contexts to learn disentangled representations without hurting the overall task accuracy. We show that task-switching networks operate in an intermediate regime between individual and parallel. In addition, we show that shared representations are produced by the emergence neurons encoding multiple tasks. Furthermore, we study the role of contexts across network processing and show its role at aligning the task with the relevant features. Finally, we investigate how the magnitude of contexts affects the performance in task-switching networks.
****
MCTransformer: Combining Transformers And Monte-Carlo Tree Search For Offline Reinforcement Learning
Gur Yaari,Lior Rokach,Rami Puzis,Gilad Katz
Recent studies explored the framing of reinforcement learning as a sequence modeling problem, and then using Transformers to generate effective solutions. In this study, we introduce MCTransformer, a framework that combines Monte-Carlo Tree Search (MCTS) with Transformers. Our approach uses an actor-critic setup, where the MCTS component is responsible for navigating previously-explored states, aided by input from the Transformer. The Transformer controls the exploration and evaluation of new states, enabling an effective and efficient evaluation of various strategies. In addition to the development of highly effective strategies, our setup enables the use of more efficient sampling compared to existing MCTS-based solutions. MCTransformer is therefore able to perform a small number of evaluations for each newly-explored node, and to do so without degrading its performance. Our evaluation, conducted on the challenging and well-known problem of SameGame, shows that our approach outperforms both Transformer-based and MCTS-based solutions.
****
One-Step Estimator for Permuted Sparse Recovery
Hang Zhang,Ping Li
This paper considers the unlabeled sparse recovery under multiple measurements, i.e., ${\mathbf{Y}} = {\mathbf{\Pi}}^{\natural}{\mathbf{X}} {\mathbf{B}}^{\natur

al} + {\mathbf{W}}$, where ${\mathbf{Y}} \in \mathbb{R}^{n\times m}, {\mathbf{\Pi}}^{\natural}\in \mathbb{R}^{n\times n}, {\mathbf{X}} \in \mathbb{R}^{n\times p}, {\mathbf{B}}^{\natural}\in \mathbb{R}^{p\times m}, {\mathbf{W}}\in \mathbb{R}^{n\times m}$ represents the observations, missing (or incomplete) correspondence information, sensing matrix, sparse signals, and additive sensing noise, respectively. Different from the previous works on multiple measurements ($m > 1$) which all focus on the sufficient samples regime, namely, $n > p$, we consider a sparse matrix $\mathbf{B}^{\natural}$ and investigate the insufficient samples regime (i.e., $n \ll p$) for the first time. To begin with, we establish the lower bound on the sample number and \emph{signal-to-noise ratio} (${\mathsf{SNR}}$) for the correct permutation recovery. Moreover, we present a simple yet effective estimator. Under mild conditions, we show that our estimator can restore the correct correspondence information with high probability. Numerical experiments are presented to corroborate our theoretical claims.

****

Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling?

Yi Tay,Mostafa Dehghani,Samira Abnar,Hyung Won Chung,William Fedus,Jinfeng Rao,Sharan Narang,Vinh Q. Tran,Dani Yogatama,Donald Metzler

There have been a lot of interest in the scaling properties of Transformer models \citep{kaplan2020scaling}. However, not much has been done on the front of investigating the effect of scaling properties of different inductive biases and model architectures. Do model architectures scale differently? If so, how does inductive bias affect scaling behaviour? How does this influence upstream (pretraining) and downstream (transfer)? This paper conducts a systematic study of scaling behaviour of ten diverse model architectures such as Transformers, Switch Transformers, Universal Transformers, Dynamic convolutions, Performers, and recently proposed MLP-Mixers. Via extensive experiments, we show that (1) architecture is an indeed an important consideration when performing scaling and (2) the best performing model can fluctuate at different scales. We believe that the findings outlined in this work has significant implications to how model architectures are currently evaluated in the community.

****

Guarded Policy Optimization with Imperfect Online Demonstrations

Zhenghai Xue,Zhenghao Peng,Quanyi Li,Zhihan Liu,Bolei Zhou

The Teacher-Student Framework (TSF) is a reinforcement learning setting where a teacher agent guards the training of a student agent by intervening and providing online demonstrations. Assuming optimal, the teacher policy has the perfect timing and capability to intervene in the learning process of the student agent, providing safety guarantee and exploration guidance. Nevertheless, in many real-world settings it is expensive or even impossible to obtain a well-performing teacher policy. In this work, we relax the assumption of a well-performing teacher and develop a new method that can incorporate arbitrary teacher policies with modest or inferior performance. We instantiate an Off-Policy Reinforcement Learning algorithm, termed Teacher-Student Shared Control (TS2C), which incorporates teacher intervention based on trajectory-based value estimation. Theoretical analysis validates that the proposed TS2C algorithm attains efficient exploration and substantial safety guarantee without being affected by the teacher's own performance. Experiments on various continuous control tasks show that our method can exploit teacher policies at different performance levels while maintaining a low training cost. Moreover, the student policy surpasses the imperfect teacher policy in terms of higher accumulated reward in held-out testing environments. Code is available at https://metadriverse.github.io/TS2C.

****

Fast Nonlinear Vector Quantile Regression

Aviv A. Rosenberg,Sanketh Vedula,Yaniv Romano,Alexander Bronstein

$$
\newcommand{\rvar}[1]{\mathrm {#1}}
\newcommand{\rvec}[1]{\boldsymbol{\mathrm{#1}}}
$$

Quantile regression (QR) is a powerful tool for estimating one or more condition

al quantiles of a target variable $\rvar{Y}$ given explanatory features $\rvec{X}$.
A limitation of QR is that it is only defined for scalar target variables, due to the formulation of its objective function, and since the notion of quantiles has no standard definition for multivariate distributions.
Recently, vector quantile regression (VQR) was proposed as an extension of QR for vector-valued target variables, thanks to a meaningful generalization of the notion of quantiles to multivariate distributions via optimal transport.
Despite its elegance, VQR is arguably not applicable in practice due to several limitations:
(i) it assumes a linear model for the quantiles of the target $\rvec{Y}$ given the features $\rvec{X}$;
(ii) its exact formulation is intractable even for modestly-sized problems in terms of target dimensions, number of regressed quantile levels, or number of features, and its relaxed dual formulation may violate the monotonicity of the estimated quantiles;
(iii) no fast or scalable solvers for VQR currently exist.

In this work we fully address these limitations, namely:
(i) We extend VQR to the non-linear case, showing substantial improvement over linear VQR;
(ii) We propose {vector monotone rearrangement}, a method which ensures the quantile functions estimated by VQR are monotone functions;
(iii) We provide fast, GPU-accelerated solvers for linear and nonlinear VQR which maintain a fixed memory footprint, and demonstrate that they scale to millions of samples and thousands of quantile levels;
(iv) We release an optimized python package of our solvers as to widespread the use of VQR in real-world applications.
****

Multi Task Learning of Different Class Label Representations for Stronger Models
Judah A Goldfeder,Boyuan Chen,Hod Lipson
We find that the way in which class labels are represented can have a powerful effect on how well models trained on them learn. In classification, the standard way of representing class labels is as one-hot vectors. We present a new way of representing class labels called Binary Labels, where each class label is a large binary vector. We further introduce a new paradigm, multi task learning on different label representations. We train a network on two tasks. The main task is to classify images based on their one-hot label, and the auxiliary task is to classify images based on their Binary Label. We show that networks trained on both tasks have many advantages, including higher accuracy across a wide variety of datasets and architectures, both when trained from scratch and when using transfer learning. Networks trained on both tasks are also much more effective when training data is limited, and seem to do especially well on more challenging problems.
****

On the Existence of a Trojaned Twin Model
Songzhu Zheng,Yikai Zhang,Weimin Lyu,Mayank Goswami,Anderson Schneider,Yuriy Nevmyvaka,Haibin Ling,Chao Chen
We study the Trojan Attack problem, where malicious attackers sabotage deep neural network models with poisoned training data. In most existing works, the effectiveness of the attack is largely overlooked; many attacks can be ineffective or inefficient for certain training schemes, e.g., adversarial training. In this paper, we adopt a novel perspective and look into the quantitative relationship between a clean model and its Trojaned counterpart. We formulate a successful attack using classic machine learning language. Under mild assumptions, we show theoretically that there exists a Trojaned model, named Trojaned Twin, that is very close to the

clean model. This attack can be achieved by simply using a universal Trojan trigger intrinsic to the data distribution. This has powerful implications in practice; the Trojaned twin model has enhanced attack efficacy and strong resiliency against detection. Empirically, we show that our method achieves consistent attack efficacy across different training schemes, including the challenging adversarial training scheme. Furthermore, this Trojaned twin model is robust against SoTA detection methods

****

On Information Maximisation in Multi-View Self-Supervised Learning

Borja Rodríguez Gálvez,Arno Blaas,Xavier Suau,Jason Ramapuram,Dan Busbridge,Luca Zappella

The strong performance of multi-view self-supervised learning (SSL) prompted the development of many different approaches (e.g. SimCLR, BYOL, and DINO). A unified understanding of how each of these methods achieves its performance has been limited by apparent differences across objectives and algorithmic details. Through the lens of information theory, we show that many of these approaches are maximising an approximate lower bound on the mutual information between the representations of multiple views of the same datum. Further, we show that this bound decomposes into a ``reconstruction" term, treated identically by all SSL methods, and an ``entropy" term, where existing SSL methods differ in their treatment. We prove that an exact optimisation of both terms of this lower bound encompasses and unifies current theoretical properties such as recovering the true latent variables of the underlying generative process (Zimmermann et al., 2021) or or isolating content from style in such true latent variables (Von Kügelgen et al., 2021). This theoretical analysis motivates a naive but principled objective (EntRec), that exactly optimises both the reconstruction and entropy terms, thus benefiting from said theoretical properties unlike other SSL frameworks. Finally, we show EntRec achieves a downstream performance on-par with existing SSL methods on ImageNet (69.7% after 400 epochs) and on an array of transfer tasks when pre-trained on ImageNet. Furthermore, EntRec is more robust to modifying the batch size, a sensitive hyperparameter in other SSL methods.

****

Leveraging Large Language Models for Multiple Choice Question Answering

Joshua Robinson,David Wingate

While large language models (LLMs) like GPT-3 have achieved impressive results on multiple choice question answering (MCQA) tasks in the zero, one, and few-shot settings, they generally lag behind the MCQA state of the art (SOTA). MCQA tasks have traditionally been presented to LLMs like cloze tasks. An LLM is conditioned on a question (without the associated answer options) and its chosen option is the one assigned the highest probability after normalization (for length, etc.). A more natural prompting approach is to present the question and answer options to the LLM jointly and have it output the symbol (e.g., “A”) associated with its chosen answer option. This approach allows the model to explicitly compare answer options, reduces computational costs, and mitigates the effects of tokenization scheme and answer option representations on answer selection. For the natural approach to be effective, the LLM it is used with must be able to associate answer options with the symbols that represent them. The LLM needs what we term multiple choice symbol binding (MCSB) ability. This ability varies greatly by model. We show that a model with high MCSB ability performs much better with the natural approach than with the traditional approach across 20 diverse datasets and largely closes the gap with the SOTA, suggesting that the MCQA ability of LLMs has been previously underestimated.

****

Learning with Logical Constraints but without Shortcut Satisfaction

Zenan Li,Zehua Liu,Yuan Yao,Jingwei Xu,Taolue Chen,Xiaoxing Ma,Jian L\"{u}

Recent studies have started to explore the integration of logical knowledge into deep learning via encoding logical constraints as an additional loss function. However, existing approaches tend to vacuously satisfy logical constraints through shortcuts, failing to fully exploit the knowledge. In this paper, we present a new framework for learning with logical constraints. Specifically, we address the shortcut satisfaction issue by introducing dual variables for logical connectives, encoding how the constraint is satisfied. We further propose a variational framework where the encoded logical constraint is expressed as a distributional loss that is compatible with the model's original training loss. The theoretical analysis shows that the proposed approach bears some nice properties, and the experimental evaluations demonstrate its superior performance in both model generalizability and constraint satisfaction.

****

Certified Training: Small Boxes are All You Need

Mark Niklas Mueller,Franziska Eckert,Marc Fischer,Martin Vechev

To obtain, deterministic guarantees of adversarial robustness, specialized training methods are used. We propose, SABR, a novel such certified training method, based on the key insight that propagating interval bounds for a small but carefully selected subset of the adversarial input region is sufficient to approximate the worst-case loss over the whole region while significantly reducing approximation errors. We show in an extensive empirical evaluation that SABR outperforms existing certified defenses in terms of both standard and certifiable accuracies across perturbation magnitudes and datasets, pointing to a new class of certified training methods promising to alleviate the robustness-accuracy trade-off.

****

Counterfactual Generation Under Confounding

Abbavaram Gowtham Reddy,Saloni Dash,Amit Sharma,Vineeth N. Balasubramanian

A machine learning model, under the influence of observed or unobserved confounders in the training data, can learn spurious correlations and fail to generalize when deployed. For image classifiers, augmenting a training dataset using counterfactual examples has been empirically shown to break spurious correlations. However, the counterfactual generation task itself becomes more difficult as the level of confounding increases. Existing methods for counterfactual generation under confounding consider a fixed set of interventions (e.g., texture, rotation) and are not flexible enough to capture diverse data-generating processes. Given a causal generative process, we formally characterize the adverse effects of confounding on any downstream tasks and show that the correlation between generative factors (attributes) can be used to quantitatively measure confounding between generative factors. To minimize such correlation, we propose a counterfactual generation method that learns to modify the value of any attribute in an image and generate new images given a set of observed attributes, even when the dataset is highly confounded. These counterfactual images are then used to regularize the downstream classifier such that the learned representations are the same across various generative factors conditioned on the class label. Our method is computationally efficient, simple to implement, and works well for any number of generative factors and confounding variables. Our experimental results on both synthetic (MNIST variants) and real-world (CelebA) datasets show the usefulness of our approach.

****

Regression with Label Differential Privacy

Badih Ghazi,Pritish Kamath,Ravi Kumar,Ethan Leeman,Pasin Manurangsi,Avinash Varadarajan,Chiyuan Zhang

We study the task of training regression models with the guarantee of _label_ differential privacy (DP). Based on a global prior distribution of label values, which could be obtained privately, we derive a label DP randomization mechanism that is optimal under a given regression loss function. We prove that the optimal mechanism takes the form of a "randomized response on bins", and propose an efficient algorithm for finding the optimal bin values. We carry out a thorough experimental evaluation on several datasets demonstrating the efficacy of our algorithm.

****

Hierarchical Abstraction for Combinatorial Generalization in Object Rearrangement

Michael Chang,Alyssa Li Dayan,Franziska Meier,Thomas L. Griffiths,Sergey Levine,Amy Zhang

Object rearrangement is a challenge for embodied agents because solving these tasks requires generalizing across a combinatorially large set of configurations of entities and their locations. Worse, the representations of these entities are unknown and must be inferred from sensory percepts. We present a hierarchical abstraction approach to uncover these underlying entities and achieve combinatorial generalization from unstructured visual inputs. By constructing a factorized transition graph over clusters of entity representations inferred from pixels, we show how to learn a correspondence between intervening on states of entities in the agent's model and acting on objects in the environment. We use this correspondence to develop a method for control that generalizes to different numbers and configurations of objects, which outperforms current offline deep RL methods when evaluated on simulated rearrangement tasks.

****

SRBGCN: Tangent space-Free Lorentz Transformations for Graph Feature Learning

Abdelrahman Mostafa,Wei Peng,Guoying Zhao

Hyperbolic graph convolutional networks have been successfully applied to represent complex graph data structures. However, optimization on Riemannian manifolds is nontrivial thus most of the existing hyperbolic networks build the network operations on the tangent space of the manifold, which is a Euclidean local approximation. This distorts the learnt features, limits the representation capacity of the network and makes it hard to optimize the network. In this work, we introduce a fully hyperbolic graph convolutional network (GCN), referred to as SRBGCN, which performs neural computations such as feature transformation and aggregation directly on the manifold, using manifold-preserving Lorentz transformations that include spatial rotation (SR) and boost (B) operations. Experiments conducted on static graph datasets for node classification and link prediction tasks validate the performance of the proposed method.

****

Transfer NAS with Meta-learned Bayesian Surrogates

Gresa Shala,Thomas Elsken,Frank Hutter,Josif Grabocka

While neural architecture search (NAS) is an intensely-researched area, approaches typically still suffer from either (i) high computational costs or (ii) lack of robustness across datasets and experiments. Furthermore, most methods start searching for an optimal architecture from scratch, ignoring prior knowledge. This is in contrast to the manual design process by researchers and engineers that leverage previous deep learning experiences by, e.g., transferring architectures from previously solved, related problems.

We propose to adopt this human design strategy and introduce a novel surrogate for NAS, that is meta-learned across prior architecture evaluations across different datasets. We utilizes Bayesian Optimization (BO) with deep-kernel Gaussian Processes, graph neural networks for the architecture embeddings and a transformer-based set encoder of datasets. As a result, our method consistently achieves state-of-the-art results on six computer vision datasets, while being as fast as one-shot NAS methods.

****

Mitigating the Limitations of Multimodal VAEs with Coordination-Based Approach

Masahiro Suzuki,Yutaka Matsuo

One of the key challenges in multimodal variational autoencoders (VAEs) is inferring a joint representation from arbitrary subsets of modalities. The state-of-the-art approach to achieving this is to sub-sample the modality subsets and learn to generate all modalities from them. However, this sub-sampling in the mixture-based approach has been shown to degrade other important features of multimodal VAEs, such as quality of generation, and furthermore, this degradation is theoretically unavoidable. In this study, we focus on another approach to learning the joint representation by bringing unimodal inferences closer to joint inference from all modalities, which does not have the above limitation. Although there have been models that can be categorized under this approach, they were derived from different backgrounds; therefore, the relation and superiority between them were not clear. To take a unified view, we first categorize them as coordination-based multimodal VAEs and show that these can be derived from the same multimodal evidence lower bound (ELBO) and that the difference in their performance is related to whether they are more tightly lower bounded. Next, we point out that these existing coordination-based models perform poorly on cross-modal generation (or cross-coherence) because they do not learn to reconstruct modalities from unimodal inferences. Therefore, we propose a novel coordination-based model that incorporates these unimodal reconstructions, which avoids the limitations of both mixture and coordination-based models. Experiments with diverse and challenging datasets show that the proposed model mitigates the limitations in multimodal VAEs and performs well in both cross-coherence and generation quality.

****

Incompatibility between Deterministic Policy and Generative Adversarial Imitation Learning

Wanying Wang,Yirui Zhou,Chaomin Shen,Yangchun Zhang,Jian Tang,Zhiyuan Xu,Yaxin Peng

Deterministic policies are widely applied in generative adversarial imitation learning (GAIL). When adopting these policies, some GAIL variants modify the reward function to avoid training instability. However, the mechanism behind this instability is still largely unknown. In this paper, we capture the instability through the underlying exploding gradients theoretically in the updating process. Our novelties lie in the following aspects: 1) By employing multivariate Gaussian policy with small covariance to approximate deterministic policy, we establish and prove the probabilistic lower bound for the exploding gradients, which can describe the degree of instability universally, while the stochastic policy will never suffer from such pathology subsequently. 2) We also prove that the modified reward function of adversarial inverse reinforcement learning (AIRL) can relieve exploding gradients, but at the expense of ``non-confidence''. Experiments and a toy demo support our analysis.

****

FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation

Sebastian Hofstätter,Jiecao Chen,Karthik Raman,Hamed Zamani

Retrieval-augmented generation models offer many benefits over standalone language models: besides a textual answer to a given query they provide provenance items retrieved from an updateable knowledge base. However, they are also more complex systems and need to handle long inputs. In this work, we introduce FiD-Light to strongly increase the efficiency of the state-of-the-art retrieval-augmented FiD model, while maintaining the same level of effectiveness. Our FiD-Light model constrains the information flow from the encoder (which encodes passages separately) to the decoder (using concatenated encoded representations). Furthermore, we adapt FiD-Light with re-ranking capabilities through textual source pointers, to improve the top-ranked provenance precision. Our experiments on a diverse set of seven knowledge intensive tasks (KILT) show FiD-Light consistently improves the Pareto frontier between query latency and effectiveness. FiD-Light with source pointing sets substantial new state-of-the-art results on six KILT tasks for combined text generation and provenance retrieval evaluation, while maintaining reasonable efficiency.

****

Contrastive Learning of Molecular Representation with Fragmented Views

Seojin Kim,Jaehyun Nam,Junsu Kim,Hankook Lee,Sungsoo Ahn,Jinwoo Shin

Molecular representation learning is a fundamental task for AI-based drug design and discovery. Contrastive learning is an attractive framework for this task, as also evidenced in various domains of representation learning, e.g., image, language, and speech. However, molecule-specific ways of constructing good positive or negative views in contrastive training under consideration of their chemical semantics have been relatively under-explored. In this paper, we consider a molecule as a bag of meaningful fragments, e.g., functional groups, by disconnecting a non-ring single bond as the semantic-preserving transformation. Then, we sug

gest to construct a complete (or incomplete) bag of fragments as the positive (or negative) views of a molecule: each fragment loses chemical substructures from the original molecule, while the union of the fragments does not. Namely, this provides easy positive and hard negative views simultaneously for contrastive representation learning so that it can selectively learn useful features and ignore nuisance features. Furthermore, we additionally suggest to optimize the torsional angle reconstruction loss around the fragmented bond to incorporate with 3D geometric structure in the pre-training dataset. Our experiments demonstrate that our scheme outperforms prior state-of-the-art molecular representation learning methods across various downstream molecule property prediction tasks.
****

Sharp Convergence Analysis of Gradient Descent for Deep Linear Neural Networks
Hongru Zhao,Jinchao Xu

This paper provides sharp rates of convergence of the gradient descent (GD) method for deep linear neural networks with different random initialization. This study touches upon one major open theoretical problem in machine learning: why deep neural networks trained with GD methods are efficient in many practical applications. While the solution of this problem is still beyond reach for general nonlinear deep neural networks, there have been extensive efforts in the literature in studying relevant questions for deep linear neural networks and there are many interesting results in this research direction. For example, recent results on the loss landscape show that even though the loss function of deep linear neural networks is non-convex, every local minimizer is also a global minimizer. When the GD method is applied to train the deep linear networks, it has been shown in the literature that the convergence behavior of the GD method depends on the initialization. In this paper, we obtain the sharp rate of convergence of GD for deep linear networks, and we show that this rate does not depend on the types of random initialization. Furthermore, we show that the depth of the network does not affect the optimal rate of convergence, provided that the width of each hidden layer is appropriately large.
****

Selective Frequency Network for Image Restoration
Yuning Cui,Yi Tao,Zhenshan Bing,Wenqi Ren,Xinwei Gao,Xiaochun Cao,Kai Huang,Alois Knoll

Image restoration aims to reconstruct the latent sharp image from its corrupted counterpart. Besides dealing with this long-standing task in the spatial domain, a few approaches seek solutions in the frequency domain in consideration of the large discrepancy between spectra of sharp/degraded image pairs. However, these works commonly utilize transformation tools, e.g., wavelet transform, to split features into several frequency parts, which is not flexible enough to select the most informative frequency component to recover. In this paper, we exploit a multi-branch and content-aware module to decompose features into separate frequency subbands dynamically and locally, and then accentuate the useful ones via channel-wise attention weights. In addition, to handle large-scale degradation blurs, we propose an extremely simple decoupling and modulation module to enlarge the receptive field via global and window-based average pooling. Integrating two developed modules into a U-Net backbone, the proposed Selective Frequency Network (SFNet) performs favorably against state-of-the-art algorithms on five image restoration tasks, including single-image defocus deblurring, image dehazing, image motion deblurring, image desnowing, and image deraining.
****

Contextualized Generative Retrieval
Hyunji Lee,JaeYoung Kim,Hoyeon Chang,Hanseok Oh,Sohee Yang,vladimir karpukhin,Yi Lu,Minjoon Seo

The text retrieval task is mainly performed in two ways: the bi-encoder approach and the generative approach. The bi-encoder approach maps the document and query embeddings to common vector space and performs a nearest neighbor search. It stably shows high performance and efficiency across different domains but has an embedding space bottleneck as it interacts in L2 or inner product space. The generative retrieval model retrieves by generating a target sequence and overcomes

the embedding space bottleneck by interacting in the parametric space. However, it fails to retrieve the information it has not seen during the training process as it depends solely on the information encoded in its own model parameters. To leverage the advantages of both approaches, we propose Contextualized Generative Retrieval model, which uses contextualized embeddings (output embeddings of a language model encoder) as vocab embeddings at the decoding step of generative retrieval. The model uses information encoded in both the non-parametric space of contextualized token embeddings and the parametric space of the generative retrieval model. Our approach of generative retrieval with contextualized vocab embeddings shows higher performance than generative retrieval with only vanilla vocab embeddings in the document retrieval task, an average of 6% higher performance in KILT (NQ, TQA) and 2X higher in NQ-320k, suggesting the benefits of using contextualized embedding in generative retrieval models.
****

Mirror Training for Input Convex Neural Network

Jiaqi Wu,Yang Weng

The input convex neural network (ICNN) aims to learn a convex function from the input to the output by using non-decreasing convex activation functions and non-negativity constraints on the weight parameters of some layers. However, in practice, it loses some representation power because of these non-negativity parameters of the hidden units, even though the design of the ``passthrough'' layer can partially address this problem. To solve issues caused by these non-negativity constraints, we use a duplication input pair trick, i.e., the negation of the original input as part of the new input in our structure. This new method will preserve the convexity of the function from the original input to the output and tackle the representation problem in training. Additionally, we design a mirror unit to address this problem further, making the network Mirror ICNN. Moreover, we propose a recurrent input convex neural network (RICNN) structure to deal with the time-series problems. The recurrent unit of the structure can be ICNN or any other convex variant of ICNN. This structure can maintain convexity by constraining the mapping from the hidden output at time step t to the input of the next time step $t+1$. The experiments can support our design, including the simple numerical curve fitting, power system hosting capacity dataset regression, and the MNIST dataset classification.
****

Scaling Up Probabilistic Circuits by Latent Variable Distillation

Anji Liu,Honghua Zhang,Guy Van den Broeck

Probabilistic Circuits (PCs) are a unified framework for tractable probabilistic models that support efficient computation of various probabilistic queries (e.g., marginal probabilities). One key challenge is to scale PCs to model large and high-dimensional real-world datasets: we observe that as the number of parameters in PCs increases, their performance immediately plateaus. This phenomenon suggests that the existing optimizers fail to exploit the full expressive power of large PCs. We propose to overcome such bottleneck by latent variable distillation: we leverage the less tractable but more expressive deep generative models to provide extra supervision over the latent variables of PCs. Specifically, we extract information from Transformer-based generative models to assign values to latent variables of PCs, providing guidance to PC optimizers. Experiments on both image and language modeling benchmarks (e.g., ImageNet and WikiText-2) show that latent variable distillation substantially boosts the performance of large PCs compared to their counterparts without latent variable distillation. In particular, on the image modeling benchmarks, PCs achieve competitive performance against some of the widely-used deep generative models, including variational autoencoders and flow-based models, opening up new avenues for tractable generative modeling. Our code can be found at https://github.com/UCLA-StarAI/LVD.
****

Oscillation Neural Ordinary Differential Equations

Muhao Guo,Yang Weng

Neural ordinary differential equations (NODEs) have received a lot of attention in recent years due to their memory efficiency. Different from traditional deep learning, it defines a continuous deep learning architecture based on the theory of ordinary differential equations (ODEs), which also improves the interpretability of deep learning. However, it has several obvious limitations, such as a NODE is not a universal approximator, it requires a large number of function evaluations (NFEs), and it has a slow convergence rate. We address these drawbacks by modeling and adding an oscillator to the framework of the NODEs. The oscillator enables the trajectories of our model to cross each other. We prove that our model is a universal approximator, even in the original input space. Due to the presence of oscillators, the flows learned by the model will be simpler, thus our model needs fewer NFEs and has a faster convergence speed. We apply our model to various tasks including classification and time series extrapolation, then compare several metrics including accuracy, NFEs, and convergence speed. The experiments show that our model can achieve better results compared to the existing baselines.

****

Improving Differentiable Neural Architecture Search by Encouraging Transferability

Parth Sheth,Pengtao Xie

Differentiable neural architecture search methods are increasingly popular due to their computational efficiency. However, these methods have unsatisfactory generalizability and stability. Their searched architectures are often degenerate with a dominant number of skip connections and perform unsatisfactorily on test data. Existing methods for solving this problem have a variety of limitations, such as cannot prevent the happening of architecture degeneration, being excessively restrictive in setting the number of skip connections, etc. To address these limitations, we propose a new approach for improving the generalizability and stability of differentiable NAS, by developing a transferability-encouraging tri-level optimization framework which improves the architecture of a main model by encouraging good transferability to an auxiliary model. Our framework involves three stages performed end-to-end: 1) train network weights of a main model; 2) transfer knowledge from the main model to an auxiliary model; 3) optimize the architecture of the main model by maximizing its transferability to the auxiliary model. We propose a new knowledge transfer approach based on matching quadruple relative similarities. Experiments on several datasets demonstrate the effectiveness of our method.

****

MA-BERT: Towards Matrix Arithmetic-only BERT Inference by Eliminating Complex Non-Linear Functions

Neo Wei Ming,Zhehui Wang,Cheng Liu,Rick Siow Mong Goh,Tao Luo

Due to their superior results, Transformer-based models such as BERT have become de facto standards in many Natural Language Processing (NLP) applications. However, the intensive use of complex non-linear functions within the Transformer architecture impairs its computing efficiency and complicates corresponding accelerator designs, because non-linear functions are generally computation-intensive and require special hardware support. In light of this, we propose MA-BERT, which allows matrix arithmetic-only operations in Transformer-based NLP models and achieves efficient inference with negligible accuracy loss. Specifically, we propose four correlated techniques that include approximating softmax with a two-layer neural network, replacing GELU with ReLU, fusing normalization layers with adjacent linear layers, and leveraging knowledge transfer from baseline models. Through these techniques, we are able to eliminate the major non-linear functions in Transformer-based models and obtain MA-BERT with only matrix arithmetic and trivial ReLU operations without compromising on accuracy. With mainly regular matrix arithmetic operations, MA-BERT enables hardware-friendly processing on various computing engines, including CPUs and GPUs. Our experimental results show that MA-BERT achieves up to 27% and 41% reduction in inference time on CPU and GPU, respectively, with comparable accuracy on many downstream tasks compared to the baseline BERT models.

****

Automatically Answering and Generating Machine Learning Final Exams

Sarah Zhang,Reece S Shuttleworth,Zad Chin,Pedro Lantigua,Saisamrit Surbehera,Gregory Hunter,Derek Austin,Yann Hicke,Leonard Tang,Sathwik Karnik,Darnell Granberry,Iddo Drori

Can a machine learn machine learning? We propose to answer this question using the same criteria we use to answer a similar question: can a human learn machine learning? We automatically answer final exams in MIT's recent large machine learning course and generate new questions at a human level. Recently, program synthesis and few-shot learning solved university-level problem set questions in mathematics and STEM courses at a human level. In this work, we solve questions from final exams that differ from problem sets in several ways: the questions are longer, have multiple parts, are more complicated, and span a broader set of topics. We provide a new dataset and benchmark of questions from machine learning final exams and code for automatically answering these questions and generating new questions. To make our dataset a reproducible benchmark, we use automatic checkers for multiple choice questions, questions with numeric answers, and questions with expression answers, and evaluate a large free language model, Meta’s OPT, and compare the results with Open AI’s GPT-3 and Codex. A student survey comparing the quality, appropriateness, and difficulty of machine-generated questions with human-written questions shows that across multiple aspects, machine-generated questions are indistinguishable from human-generated questions and are suitable for final exams. We perform ablation studies comparing zero-shot learning with few-shot learning, chain-of-thought prompting, GPT-3 and OPT pre-trained on text and Codex fine-tuned on code on a range of machine learning topics and find that few-shot learning methods perform best. We make our data and code publicly available for the machine learning community.

****

CAT: Collaborative Adversarial Training

xingbin liu,Huafeng Kuang,Xianming Lin,GUANNAN JIANG,YONGJIAN WU,Xi Wang,Rongrong Ji

Adversarial training can improve the robustness of neural networks. Previous adversarial training methods focus on a single training strategy and do not consider the collaboration between different training strategies. In this paper, we find different adversarial training methods have distinct robustness for sample instances. For example, an instance can be correctly classified by a model trained using standard adversarial training (AT) but not by a model trained using TRADES, and vice versa. Based on this phenomenon, we propose a collaborative adversarial training framework to improve the robustness of neural networks. Specifically, we simultaneously use different adversarial training methods to train two robust models from scratch. We input the adversarial examples generated by each network to the peer network and use the logit of the peer network to guide the training of its network. Collaborative Adversarial Training (CAT) can improve both robustness and accuracy. Finally, Extensive experiments on CIFAR-10 and CIFAR-100 validated the effectiveness of our method.

CAT achieved new state-of-the-art robustness without using any additional data on CIFAR-10 under the Auto-Attack benchmark.

****

Efficient Certified Training and Robustness Verification of Neural ODEs

Mustafa Zeqiri,Mark Niklas Mueller,Marc Fischer,Martin Vechev

Neural Ordinary Differential Equations (NODEs) are a novel neural architecture, built around initial value problems with learned dynamics which are solved during inference. Thought to be inherently more robust against adversarial perturbations, they were recently shown to be vulnerable to strong adversarial attacks, highlighting the need for formal guarantees. However, despite significant progress in robustness verification for standard feed-forward architectures, the verification of high dimensional NODEs remains an open problem. In this work we address this challenge and propose GAINS, an analysis framework for NODEs combining three key ideas: (i) a novel class of ODE solvers, based on variable but discrete time steps, (ii) an efficient graph representation of solver trajectories, and (iii) a novel abstraction algorithm operating on this graph representation. Together, these advances enable the efficient analysis and certified training of high

-dimensional NODEs, by reducing the runtime from an intractable $\mathcal{O}(\exp(d)+\exp(T))$ to $\mathcal{O}(d+T^2\log^2T)$ in the dimensionality d and integration time T. In an extensive evaluation on computer vision (MNIST and Fashion-MNIST) and time-series forecasting (Physio-Net) problems, we demonstrate the effectiveness of both our certified training and verification methods.

****

Arbitrary Virtual Try-on Network: Characteristics Representation and Trade-off between Body and Clothing

Yu Liu,Mingbo Zhao,Zhao Zhang,Jicong Fan,Yang Lou,Shuicheng Yan

Deep learning based virtual try-on system has achieved some encouraging progress recently, but there still remain several big challenges that need to be solved, such as trying on arbitrary clothes of all types, trying on the clothes from one category to another and generating image-realistic results with few artifacts. To handle this issue, we propose the Arbitrary Virtual Try-On Network (AVTON) that is utilized for all-type clothes, which can synthesize realistic try-on images by preserving and trading off characteristics of the target clothes and the reference person. Our approach includes three modules: 1) Limbs Prediction Module, which is utilized for predicting the human body parts by preserving the characteristics of the reference person. This is especially good for handling cross-category try-on task (e.g., long sleeves \(\leftrightarrow\) short sleeves or long pants \(\leftrightarrow\) skirts, etc.), where the exposed arms or legs with the skin colors and details can be reasonably predicted; 2) Improved Geometric Matching Module, which is designed to warp clothes according to the geometry of the target person. We improve the TPS-based warping method with a compactly supported radial function (Wendland's \(\Psi\)-function); 3) Trade-Off Fusion Module, which is to trade off the characteristics of the warped clothes and the reference person. This module is to make the generated try-on images look more natural and realistic based on a fine-tuning symmetry of the network structure. Extensive simulations are conducted and our approach can achieve better performance compared with the state-of-the-art virtual try-on methods.

****

A Benchmark Dataset for Learning from Label Proportions

Anand Paresh Brahmbhatt,Mohith Pokala,Rishi Saket,Aravindan Raghuveer

Learning from label proportions (LLP) has recently emerged as an important technique of weakly supervised learning on aggregated labels. In LLP, a model is trained on groups (a.k.a bags) of feature-vectors and their corresponding label proportions to predict labels for individual feature-vectors. While previous works have developed a variety of techniques for LLP, including novel loss functions, model architectures and their optimization, they typically evaluated their methods on pseudo-synthetically generated LLP training data using common small scale supervised learning datasets by randomly sampling or partitioning their instances into bags. Despite growing interest in this important task there are no large scale open source LLP benchmarks to compare various approaches. Construction of such a benchmark is hurdled by two challenges a) lack of natural large scale LLP like data, b) large number of mostly artificial methods of forming bags from instance level datasets.

In this paper we propose LLP-Bench: a large scale LLP benchmark constructed from the Criteo Kaggle CTR dataset. We do an in-depth, systematic study of the Criteo dataset and propose a methodology to create a benchmark as a collection of diverse and large scale LLP datasets. We choose the Criteo dataset since it admits multiple natural collections of bags formed by grouping subsets of its 26 categorical features. We analyze all bag collections obtained through grouping by one or two categorical features, in terms of their bag-level statistics as well as embedding based distance metrics quantifying the geometric separation of bags. We then propose to include in LLP-Bench a few groupings to fairly represent real world bag distributions.

We also measure the performance of state of the art models, loss functions (adapted to LLP) and optimizers on LLP-Bench. We perform a series of ablations and explain the performance of various techniques on LLP-Bench. To the best of our knowledge LLP-Bench is the first open source benchmark for the LLP task. We hope t

hat the proposed benchmark and the evaluation methodology will be used by ML researchers and practitioners to better understand and hence devise state of art LLP algorithms.
****

UL2: Unifying Language Learning Paradigms

Yi Tay,Mostafa Dehghani,Vinh Q. Tran,Xavier Garcia,Jason Wei,Xuezhi Wang,Hyung Won Chung,Dara Bahri,Tal Schuster,Steven Zheng,Denny Zhou,Neil Houlsby,Donald Metzler

Existing pre-trained models are generally geared towards a particular class of problems. To date, there seems to be still no consensus on what the right architecture and pre-training setup should be. This paper presents a unified framework for pre-training models that are universally effective across datasets and setups. We begin by disentangling architectural archetypes with pre-training objectives -- two concepts that are commonly conflated. Next, we present a generalized and unified perspective for self-supervision in NLP and show how different pre-training objectives can be cast as one another and how interpolating between different objectives can be effective. We then propose Mixture-of-Denoisers (MoD), a pre-training objective that combines diverse pre-training paradigms together. We furthermore introduce a notion of mode switching, wherein downstream fine-tuning is associated with specific pre-training schemes. We conduct extensive ablative experiments to compare multiple pre-training objectives and find that our method pushes the Pareto-frontier by outperforming T5 and/or GPT-like models across multiple diverse setups. Finally, by scaling our model up to 20B parameters, we achieve SOTA performance on 50 well-established supervised NLP tasks ranging from language generation (with automated and human evaluation), language understanding, text classification, question answering, commonsense reasoning, long text reasoning, structured knowledge grounding and information retrieval. Our model also achieve strong results at in-context learning, outperforming 175B GPT-3 on zero-shot SuperGLUE and tripling the performance of T5-XXL on one-shot summarization. Finally, we show that UL2 20B works well with chain-of-thought prompting and reasoning, making it an appealing choice for research into reasoning at a small to medium scale of 20B parameters. We release Flax-based T5X model checkpoints for the 20B model publicly.

****

Emergence of Exploration in Policy Gradient Reinforcement Learning via Resetting

Sotetsu Koyamada,Paavo Parmas,Tadashi Kozuno,Shin Ishii

In reinforcement learning (RL), many exploration methods explicitly promote stochastic policies, e.g., by adding an entropy bonus. We argue that exploration only matters in RL because the agent repeatedly encounters the same or similar states, so that it is beneficial to gradually improve the performance over the encounters; otherwise, the greedy policy would be optimal. Based on this intuition, we propose ReMax, an objective for RL whereby stochastic exploration arises as an emergent property, without adding any explicit exploration bonus. In ReMax, an episode is modified so that the agent can reset to previous states in the trajectory, and the agent’s goal is to maximize the best return in the trajectory tree. We show that this ReMax objective can be directly optimized with an unbiased policy gradient method. Experiments confirm that ReMax leads to the emergence of a stochastic exploration policy, and improves the performance compared to RL with no exploration bonus.
****

CASR: Generating Complex Sequences with Autoregressive Self-Boost Refinement

Hongwei Han,Mengyu Zhou,Shi Han,Xiu Li,Dongmei Zhang

There are sequence generation tasks where the best order to generate the target sequence is not left-to-right. For example, an answer to the Sudoku game, a structured code like s-expression, and even a logical natural language answer where the analysis may be generated after the decision. We define the target sequences of those tasks as complex sequences. Obviously, a complex sequence should be constructed with multiple logical steps, and has dependencies among each part of itself (e.g. decisions depend on analyses). It's a great challenge for the classi c left-to-right autoregressive generation system to generate complex sequences. Current approaches improve one-pass left-to-right generation on NLG tasks by generating different heuristic intermediate sequences in multiple stages. However, for complex sequences, the heuristic rules to break down them may hurt performance, and increase additional exposure bias. To tackle these challenges, we propose a PLM-friendly autoregressive self-boost refinement framework, CASR. When training, CASR inputs the predictions generated by the model itself at the previous refinement step (instead of those produced by heuristic rules). To find an optimal design, we also discuss model architecture, parameter efficiency and initialization strategy. By evaluating CASR on Sudoku, WebQSP, MTOP and KVRET through controlled experiments and empirical studies, we find that CASR produces high-quality outputs. CASR also improves Accuracy on Sudoku (70.93% --> 97.28%) and achieves state-of-the-art performance on KVRET with Micro F1 score (67.88% --> 70.00%).

****

SciRepEval: A Multi-Format Benchmark for Scientific Document Representations
Amanpreet Singh,Mike D'Arcy,Arman Cohan,Doug Downey,Sergey Feldman

Learned representations of scientific documents can serve as valuable input features for downstream tasks, without the need for further fine-tuning. However, existing benchmarks for evaluating these representations fail to capture the diversity of relevant tasks. In response, we introduce SciRepEval, the first comprehensive benchmark for training and evaluating scientific document representations. It includes 25 challenging and realistic tasks across four formats: classification, regression, ranking and search. We then use the benchmark to study and improve the generalization ability of scientific document representation models. We show how state-of-the-art models struggle to generalize across task formats, and that simple multi-task training fails to improve them. However, a new approach that learns multiple embeddings per document, each tailored to a different task format, can improve performance.

We experiment with task-format-specific control codes and adapters in a multi-task setting and find that they outperform the existing single-embedding state-of-the-art by up to 1.5 points absolute.

****

On the convergence of SGD under the over-parameter setting
ruinan Jin,Yiwei Wang,Wei Liu,Baoxiang Wang

With the improvement of computing power, over-parameterized models get increasingly popular in machine learning. This type of model is usually with a complicated, non-smooth, and non-convex loss function landscape. However, when we train the model, simply using the first-order optimization algorithm like stochastic gradient descent (SGD) could acquire some good results, in both training and testing, albeit that SGD is known to not guarantee convergence for non-smooth and non-convex cases. Theoretically, it was previously proved that in training, SGD converges to the global optimum with probability $1 - \epsilon$, but only for certain models and ϵ depends on the model complexity. It was also observed that SGD tends to choose a flat minimum, which preserves its training performance in testing. In this paper, we first prove that SGD could iterate to the global optimum almost surely under arbitrary initial value and some mild assumptions on the loss function. Then, we prove that if the learning rate is larger than a value depending on the structure of a global minimum, the probability of converging to this global optimum is zero. Finally, we acquire the asymptotic convergence rate based on the local structure of the global optimum.

****

MASTER: Multi-task Pre-trained Bottlenecked Masked Autoencoders are Better Dense Retrievers
Kun Zhou,Xiao Liu,Yeyun Gong,Xin Zhao,Daxin Jiang,Nan Duan,Ji-Rong Wen

Dense retrieval aims to map queries and passages into low-dimensional vector space for efficient similarity measuring, showing promising effectiveness in various large-scale retrieval tasks. Since most existing methods commonly adopt pre-trained Transformers (\eg BERT) for parameter initialization, some work focuses on proposing new pre-training tasks for compressing the useful semantic informatio n from passages into dense vectors, achieving remarkable performances. However, it is still challenging to effectively capture the rich semantic information and relations about passages into the dense vectors via one single particular pre-training task. In this work, we propose a multi-task pre-trained model, MASTER, that unifies and integrates multiple pre-training tasks with different learning objectives under the bottlenecked masked autoencoder architecture. Concretely, MASTER utilizes a multi-decoder architecture to integrate three types of pre-training tasks: corrupted passages recovering, related passage recovering and PLMs outputs recovering. By incorporating a shared deep encoder, we construct a representation bottleneck in our architecture, compressing the abundant semantic information across tasks into dense vectors. The first two types of tasks concentrate on the semantic information of passages and capturing relationships among them within the pre-training corpus. The third can capture the knowledge beyond the corpus from external PLMs (\eg GPT-2). Extensive experiments on several large-scale passage retrieval datasets have shown that our approach outperforms the previous state-of-the-art dense retrieval methods.

****

Bitrate-Constrained DRO: Beyond Worst Case Robustness To Unknown Group Shifts

Amrith Setlur,Don Dennis,Benjamin Eysenbach,Aditi Raghunathan,Chelsea Finn,Virginia Smith,Sergey Levine

Training machine learning models robust to distribution shifts is critical for real-world applications. Some robust training algorithms (e.g., Group DRO) specialize to group shifts and require group information on all training points. Other methods (e.g., CVaR DRO) that do not need group annotations can be overly conservative, since they naively upweight high loss points which may form a contrived set that does not correspond to any meaningful group in the real world (e.g., when the high loss points are randomly mislabeled training points). In this work, we address limitations in prior approaches by assuming a more nuanced form of group shift: conditioned on the label, we assume that the true group function (indicator over group) is simple. For example, we may expect that group shifts occur along low bitrate features (e.g., image background, lighting). Thus, we aim to learn a model that maintains high accuracy on simple group functions realized by these low bitrate features, that need not spend valuable model capacity achieving high accuracy on contrived groups of examples. Based on this, we consider the two-player game formulation of DRO where the adversary's capacity is bitrate-constrained. Our resulting practical algorithm, Bitrate-Constrained DRO (\bdro), does not require group information on training samples yet matches the performance of Group DRO on datasets that have training group annotations and that of CVaR DRO on long-tailed distributions. Our theoretical analysis reveals that in some settings \bdro objective can provably yield statistically efficient and less conservative solutions than unconstrained CVaR DRO.

****

Some Practical Concerns and Solutions for Using Pretrained Representation in Industrial Systems

Da Xu

Deep learning has dramatically changed the way data scientists and engineers craft features -- the once tedious process of measuring and constructing can now be achieved by training learnable representations. Recent work shows pretraining can endow representations with relevant signals, and in practice they are often used as feature vectors in downstream models. In real-world production, however, we have encountered key problems that cannot be justified by existing knowledge. They raise concerns that the naive use of pretrained representation as feature vector could lead to unwarranted and suboptimal solution.

Our investigation reveals critical insights into the gap of uniform convergence for analyzing pretrained representations, their stochastic nature under gradient descent optimization, what does model convergence means to them, and how they might interact with downstream tasks. Inspired by our analysis, we explore a simple yet powerful approach that can refine pretrained representation in multiple ways, which we call "Featurizing Pretrained Representations". Our work balances practicality and rigor, and contributes to both applied and theoretical research of representation learning.

****

Exphormer: Scaling Graph Transformers with Expander Graphs

Hamed Shirzad,Ameya Velingker,Balaji Venkatachalam,Danica J. Sutherland,Ali Kemal Sinop

Graph transformers have emerged as a promising architecture for a variety of graph learning and representation tasks. Despite their successes, it remains challenging to scale graph transformers to large graphs while maintaining accuracy competitive with message-passing networks. In this paper, we introduce Exphormer, a framework for building powerful and scalable graph transformers. Exphormer consists of a sparse attention mechanism based on expander graphs, whose mathematical characteristics, such as spectral expansion, and sparsity, yield graph transformers with complexity only linear in the size of the graph, while allowing us to prove desirable theoretical properties of the resulting transformer models. We show that incorporating Exphormer into the recently-proposed GraphGPS framework produces models with competitive empirical results on a wide variety of graph datasets, including state-of-the-art results on three datasets. We also show that Exphormer can scale to datasets on larger graphs than shown in previous graph transformer architectures.

****

Feature selection and low test error in shallow low-rotation ReLU networks

Matus Telgarsky

This work establishes low test error of gradient flow (GF) and stochastic gradient descent (SGD) on two-layer ReLU networks with standard initialization scale, in three regimes where key sets of weights rotate little (either naturally due to GF and SGD, or due to an artificial constraint), and making use of margins as the core analysis technique. The first regime is near initialization, specifically until the weights have moved by $\mathcal{O}(\sqrt m)$, where m denotes the network width, which is in sharp contrast to the $\mathcal{O}(1)$ weight motion allowed by the Neural Tangent Kernel (NTK); here it is shown that GF and SGD only need a network width and number of samples inversely proportional to the NTK margin, and moreover that GF attains at least the NTK margin itself and in particular escapes bad KKT points of the margin objective, whereas prior work could only establish nondecreasing but arbitrarily small margins. The second regime is the Neural Collapse (NC) setting, where data lies in well-separated groups, and the sample complexity scales with the number of groups; here the contribution over prior work is an analysis of the entire GF trajectory from initialization. Lastly, if the inner layer weights are constrained to change in norm only and cannot rotate, then GF with large widths achieves globally maximal margins, and its sample complexity scales with their inverse; this is in contrast to prior work, which required infinite width and a tricky dual convergence assumption.

****

Backpropagation through Combinatorial Algorithms: Identity with Projection Works

Subham Sekhar Sahoo,Anselm Paulus,Marin Vlastelica,Vít Musil,Volodymyr Kuleshov,Georg Martius

Embedding discrete solvers as differentiable layers has given modern deep learning architectures combinatorial expressivity and discrete reasoning capabilities. The derivative of these solvers is zero or undefined, therefore a meaningful replacement is crucial for effective gradient-based learning. Prior works rely on smoothing the solver with input perturbations, relaxing the solver to continuous problems, or interpolating the loss landscape with techniques that typically require additional solver calls, introduce extra hyper-parameters, or compromise performance. We propose a principled approach to exploit the geometry of the discrete solution space to treat the solver as a negative identity on the backward pass and further provide a theoretical justification. Our experiments demonstrate that such a straightforward hyper-parameter-free approach is able to compete with previous more complex methods on numerous experiments such as backpropagation through discrete samplers, deep graph matching, and image retrieval. Furthermore, we substitute the previously proposed problem-specific and label-dependent ma rgin with a generic regularization procedure that prevents cost collapse and increases robustness.

`*************************************************`

Coupled Multiwavelet Operator Learning for Coupled Differential Equations

Xiongye Xiao,Defu Cao,Ruochen Yang,Gaurav Gupta,Gengshuo Liu,Chenzhong Yin,Radu Balan,Paul Bogdan

Coupled partial differential equations (PDEs) are key tasks in modeling the complex dynamics of many physical processes. Recently, neural operators have shown the ability to solve PDEs by learning the integral kernel directly in Fourier/Wavelet space, so the difficulty of solving the coupled PDEs depends on dealing with the coupled mappings between the functions. Towards this end, we propose a \textit{coupled multiwavelets neural operator} (CMWNO) learning scheme by decoupling the coupled integral kernels during the multiwavelet decomposition and reconstruction procedures in the Wavelet space. The proposed model achieves significantly higher accuracy compared to previous learning-based solvers in solving the coupled PDEs including Gray-Scott (GS) equations and the non-local mean field game (MFG) problem. According to our experimental results, the proposed model exhibits a $2X-4X$ improvement relative $L$2 error compared to the best results from the state-of-the-art models.

`*************************************************`

Mid-Vision Feedback

Michael Maynord,Eadom T Dessalene,Cornelia Fermuller,Yiannis Aloimonos

Feedback plays a prominent role in biological vision, where perception is modulated based on agents' evolving expectations and world model. We introduce a novel mechanism which modulates perception based on high level categorical expectations: Mid-Vision Feedback (MVF). MVF associates high level contexts with linear transformations. When a context is "expected" its associated linear transformation is applied over feature vectors in a mid level of a network. The result is that mid-level network representations are biased towards conformance with high level expectations, improving overall accuracy and contextual consistency. Additionally, during training mid-level feature vectors are biased through introduction of a loss term which increases the distance between feature vectors associated with different contexts. MVF is agnostic as to the source of contextual expectations, and can serve as a mechanism for top down integration of symbolic systems with deep vision architectures. We show the superior performance of MVF to post-hoc filtering for incorporation of contextual knowledge, and show superior performance of configurations using predicted context (when no context is known a priori) over configurations with no context awareness.

`*************************************************`

Cross-Window Self-Training via Context Variations from Sparsely-Labeled Time Series

Yooju Shin,Susik Yoon,Hwanjun Song,Jae-Gil Lee,Byung Suk Lee

A real-world time series is often sparsely labeled due to the expensive annotation cost. Recently, self-training methods have been applied to a dataset with few labels to infer the labels of unlabeled augmented instances. Accelerating this trend for time-series data, fully taking advantage of its sequential nature, we propose a novel data augmentation approach called context-additive augmentation, which allows a target instance to be augmented easily by adding preceding and succeeding instances to form an augmented instance. Unlike the existing augmentation techniques which may alter the target instance by directly perturbing its features, it preserves a target instance as is but still gives various augmented instances with varying contexts. Additionally, we propose a cross-window self-training framework based on the context-additive augmentation. The framework first augments target instances by applying context-varying windows over a given time series. Then, the framework derives reliability-based cross-window labels and uses them to maintain consistency among augmented instances across the windows. Extensive experiments using real datasets show that the framework outperforms the existing state-of-the-art self-training methods.

`*************************************************`

Safe Reinforcement Learning From Pixels Using a Stochastic Latent Representation

Yannick Hogewind,Thiago D. Simão,Tal Kachman,Nils Jansen

We address the problem of safe reinforcement learning from pixel observations. Inherent challenges in such settings are (1) a trade-off between reward optimization and adhering to safety constraints, (2) partial observability, and (3) high-dimensional observations. We formalize the problem in a constrained, partially observable Markov decision process framework, where an agent obtains distinct reward and safety signals. To address the curse of dimensionality, we employ a novel safety critic using the stochastic latent actor-critic (SLAC) approach. The latent variable model predicts rewards and safety violations, and we use the safety critic to train safe policies. Using well-known benchmark environments, we demonstrate competitive performance over existing approaches regarding computational requirements, final reward return, and satisfying the safety constraints.

****

TrojText: Test-time Invisible Textual Trojan Insertion

Yepeng Liu,Bo Feng,Qian Lou

In Natural Language Processing (NLP), intelligent neuron models can be susceptible to textual Trojan attacks. Such attacks occur when Trojan models behave normally for standard inputs but generate malicious output for inputs that contain a specific trigger. Syntactic-structure triggers, which are invisible, are becoming more popular for Trojan attacks because they are difficult to detect and defend against. However, these types of attacks require a large corpus of training data to generate poisoned samples with the necessary syntactic structures for Trojan insertion. Obtaining such data can be difficult for attackers, and the process of generating syntactic poisoned triggers and inserting Trojans can be time-consuming. This paper proposes a solution called TrojText, which aims to determine whether invisible textual Trojan attacks can be performed more efficiently and cost-effectively without training data. The proposed approach, called the Representation-Logit Trojan Insertion (RLI) algorithm, uses smaller sampled test data instead of large training data to achieve the desired attack. The paper also introduces two additional techniques, namely the accumulated gradient ranking (AGR) and Trojan Weights Pruning (TWP), to reduce the number of tuned parameters and the attack overhead. The TrojText approach was evaluated on three datasets (AG’s News, SST-2, and OLID) using three NLP models (BERT, XLNet, and DeBERTa). The experiments demonstrated that the TrojText approach achieved a 98.35% classification accuracy for test sentences in the target class on the BERT model for the AG’s News dataset. The source code for TrojText is available at https://github.com/UCF-ML-Research/TrojText.

****

An Experiment Design Paradigm using Joint Feature Selection and Task Optimization

Stefano B Blumberg,Hongxiang Lin,Yukun Zhou,Paddy John Slator,Daniel C. Alexander

This paper presents a subsampling-task paradigm for data-driven task-specific experiment design (ED) and a novel method in populationwide supervised feature selection (FS). Optimal ED, the choice of sampling points under constraints of limited acquisition-time, arises in a wide variety of scientific and engineering contexts. However the continuous optimization used in classical approaches depend on a-priori parameter choices and challenging non-convex optimization landscapes. This paper proposes to replace this strategy with a subsampling-task paradigm, analogous to populationwide supervised FS. In particular, we introduce JOFSTO, which performs JOint Feature Selection and Task Optimization. JOFSTO jointly optimizes two coupled networks: one for feature scoring, which provides the ED, the other for execution of a downstream task or process. Unlike most FS problems, e.g. selecting protein expressions for classification, ED problems typically select from highly correlated globally informative candidates rather than seeking a small number of highly informative features among many uninformative features. JOFSTO's construction efficiently identifies potentially correlated, but effective subsets and returns a trained task network. We demonstrate the approach using parameter estimation and mapping problems in quantitative MRI, where economical ED is crucial for clinical application. Results from simulations and empi rical data show the subsampling-task paradigm strongly outperforms classical ED, and within our paradigm, JOFSTO outperforms state-of-the-art supervised FS techniques. JOFSTO extends immediately to wider image-based ED problems and other scenarios where the design must be specified globally across large numbers of acquisitions. Our code is available for reviewers https://www.dropbox.com/scl/fo/qe6vb1w6fuf869hx4ht0k/h?dl=0&rlkey=og8czcorurl57jbiixio7hcjt

****

Multi-Objective Online Learning

Jiyan Jiang,Wenpeng Zhang,Shiji Zhou,Lihong Gu,Xiaodong Zeng,Wenwu Zhu

This paper presents a systematic study of multi-objective online learning. We first formulate the framework of Multi-Objective Online Convex Optimization, which encompasses a novel multi-objective regret. This regret is built upon a sequence-wise extension of the commonly used discrepancy metric Pareto suboptimality gap in zero-order multi-objective bandits. We then derive an equivalent form of the regret, making it amenable to be optimized via first-order iterative methods. To motivate the algorithm design, we give an explicit example in which equipping OMD with the vanilla min-norm solver for gradient composition will incur a linear regret, which shows that merely regularizing the iterates, as in single-objective online learning, is not enough to guarantee sublinear regrets in the multi-objective setting. To resolve this issue, we propose a novel min-regularized-norm solver that regularizes the composite weights. Combining min-regularized-norm with OMD results in the Doubly Regularized Online Mirror Multiple Descent algorithm. We further derive the multi-objective regret bound for the proposed algorithm, which matches the optimal bound in the single-objective setting. Extensive experiments on real-world datasets verify the effectiveness of the proposed algorithm.

****

Improved Training of Physics-Informed Neural Networks Using Energy-Based Priors: a Study on Electrical Impedance Tomography

Akarsh Pokkunuru,Pedram Rooshenas,Thilo Strauss,Anuj Abhishek,Taufiquar Khan

Physics-informed neural networks (PINNs) are attracting significant attention for solving partial differential equation (PDE) based inverse problems, including electrical impedance tomography (EIT). EIT is non-linear and especially its inverse problem is highly ill-posed. Therefore, successful training of PINN is extremely sensitive to interplay between different loss terms and hyper-parameters, including the learning rate. In this work, we propose a Bayesian approach through data-driven energy-based model (EBM) as a prior, to improve the overall accuracy and quality of tomographic reconstruction. In particular, the EBM is trained over the possible solutions of the PDEs with different boundary conditions. By imparting such prior onto physics-based training, PINN convergence is expedited by more than ten times faster to the PDE’s solution. Evaluation outcome shows that our proposed method is more robust for solving the EIT problem. Our code is available at: https://rooshenasgroup.github.io/eit_ebprior.

****

Efficient Bayesian Optimization with Deep Kernel Learning and Transformer Pre-trained on Muliple Heterogeneous Datasets

Wenlong Lyu,Shoubo Hu,Jie Chuai,Zhitang Chen

Bayesian optimization (BO) is widely adopted in black-box optimization problems and it relies on a surrogate model to approximate the black-box response function. With the increasing number of black-box optimization tasks solved and even more to solve, the ability to learn from multiple prior tasks to jointly pre-train a surrogate model is long-awaited to further boost optimization efficiency. In this paper, we propose a simple approach to pre-train a surrogate, which is a Gaussian process (GP) with a kernel defined on deep features learned from a Transformer-based encoder, using datasets from prior tasks with possibly heterogeneous input spaces. In addition, we provide a simple yet effective mix-up initialization strategy for input tokens corresponding to unseen input variables and therefore accelerate new tasks' convergence. Experiments on both synthetic and real benchmark problems demonstrate the effectiveness of our proposed pre-training and transfer BO strategy over existing methods.

****

Robustness Guarantees for Adversarially Trained Neural Networks

Poorya Mianjy,Raman Arora

We study robust adversarial training of two-layer neural networks with Leaky ReLU activation function as a bi-level optimization problem. In particular, for the inner-loop that implements the PGD attack, we propose maximizing a lower bound on the 0/1-loss by reflecting a surrogate loss about the origin. This allows us to give a convergence guarantee for the inner-loop PGD attack and precise iteration complexity results for end-to-end adversarial training, which hold for any width and initialization in a realizable setting.

****

Fast-PINN for Complex Geometry: Solving PDEs with Boundary Connectivity Loss

Jian Cheng Wong,Pao-Hsiung Chiu,Chin Chun Ooi,My Ha Dao,Yew-Soon Ong

We present a novel loss formulation for efficient learning of complex dynamics from governing physics, typically described by partial differential equations (PDEs), using physics-informed neural networks (PINNs). In our experiments, existing versions of PINNs are seen to learn poorly in many problems, especially for complex geometries, as it becomes increasingly difficult to establish appropriate sampling strategy at the near boundary region. Overly dense sampling can adversely impede training convergence if the local gradient behaviors are too complex to be adequately modelled by PINNs. On the other hand, if the samples are too sparse, PINNs may over-fit the near boundary region, leading to incorrect solution. To prevent such issues, we propose a new Boundary Connectivity (BCXN) loss function which provides local structure approximation at the boundary. Our BCXN-loss can implicitly or explicitly impose such approximations during training, thus facilitating fast physics-informed learning across entire problem domains with order of magnitude fewer training samples. This method shows a few orders of magnitude smaller errors than existing methods in terms of the standard L2-norm metric, while using dramatically fewer training samples and iterations. Our proposed Fast-PINN method does not pose any requirement on the differentiable property of the networks, and we demonstrate its benefits and ease of implementation on both multi-layer perceptron and convolutional neural network versions as commonly used in current physics-informed neural network literature.

****

Noise Transforms Feed-Forward Networks into Sparse Coding Networks

Trenton Bricken,Bruno Olshausen,Gabriel Kreiman

A hallmark of biological neural networks, which distinguishes them from their artificial counterparts, is the high degree of sparsity in their activations. Here, we show that by simply injecting symmetric, random, noise during training in reconstruction or classification tasks, artificial neural networks with ReLU activation functions eliminate this difference; the neurons converge to a sparse coding solution where only a small fraction are active for any input. The resulting network learns receptive fields like those of primary visual cortex and remains sparse even when noise is removed in later stages of learning.

****

DEFENDING BACKDOOR ATTACKS VIA ROBUSTNESS AGAINST NOISY LABEL

Boyang Liu,Zhuangdi Zhu,Yijiang Pang,Pang-Ning Tan,Jiayu Zhou

Many deep neural networks are vulnerable to backdoor poisoning attacks, in which an adversary strategically injects a backdoor trigger into a small fraction of the training data. The trigger can later be applied during inference to manipulate prediction labels. While the data label could be changed to arbitrary values by an adversary, the extent of corruption injected into the feature values is strictly limited to keep the backdoor attack in disguise, which leads to a resemblance between the backdoor attack and a milder attack that involves only noisy labels.

This paper investigates an intriguing question: \textit{Can we leverage algorithms that defend against noisy label corruptions to defend against general backdoor attacks?} We first discuss the limitations of directly using current noisy-label defense algorithms to defend against backdoor attacks. We then propose a meta-algorithm for both supervised and semi-supervised settings that transforms an e xisting noisy label defense algorithm into one that protects against backdoor attacks. Extensive experiments on different settings show that, by introducing a lightweight alteration for minimax optimization to the existing noisy-label defense algorithms, the robustness against backdoor attacks can be substantially improved, while the initial form of those algorithms would fail in the presence of a backdoor attack.

****

A Kernel Perspective of Skip Connections in Convolutional Networks

Daniel Barzilai,Amnon Geifman,Meirav Galun,Ronen Basri

Over-parameterized residual networks (ResNets) are amongst the most successful convolutional neural architectures for image processing. Here we study their properties through their Gaussian Process and Neural Tangent kernels. We derive explicit formulas for these kernels, analyze their spectra, and provide bounds on their implied condition numbers. Our results indicate that (1) with ReLU activation, the eigenvalues of these residual kernels decay polynomially at a similar rate compared to the same kernels when skip connections are not used, thus maintaining a similar frequency bias; (2) however, residual kernels are more locally biased. Our analysis further shows that the matrices obtained by these residual kernels yield favorable condition numbers at finite depths than those obtained without the skip connections, enabling therefore faster convergence of training with gradient descent.

****

Ordered GNN: Ordering Message Passing to Deal with Heterophily and Over-smoothing

Yunchong Song,Chenghu Zhou,Xinbing Wang,Zhouhan Lin

Most graph neural networks follow the message passing mechanism. However, it faces the over-smoothing problem when multiple times of message passing is applied to a graph, causing indistinguishable node representations and prevents the model to effectively learn dependencies between farther-away nodes. On the other hand, features of neighboring nodes with different labels are likely to be falsely mixed, resulting in the heterophily problem. In this work, we propose to order the messages passing into the node representation, with specific blocks of neurons targeted for message passing within specific hops. This is achieved by aligning the hierarchy of the rooted-tree of a central node with the ordered neurons in its node representation. Experimental results on an extensive set of datasets show that our model can simultaneously achieve the state-of-the-art in both homophily and heterophily settings, without any targeted design. Moreover, its performance maintains pretty well while the model becomes really deep, effectively preventing the over-smoothing problem. Finally, visualizing the gating vectors shows that our model learns to behave differently between homophily and heterophily settings, providing an explainable graph neural model.

****

Sparse Distributed Memory is a Continual Learner

Trenton Bricken,Xander Davies,Deepak Singh,Dmitry Krotov,Gabriel Kreiman

Continual learning is a problem for artificial neural networks that their biological counterparts are adept at solving. Building on work using Sparse Distributed Memory (SDM) to connect a core neural circuit with the powerful Transformer model, we create a modified Multi-Layered Perceptron (MLP) that is a strong continual learner. We find that every component of our MLP variant translated from biology is necessary for continual learning. Our solution is also free from any memory replay or task information, and introduces novel methods to train sparse networks that may be broadly applicable.

****

Optimistic Exploration in Reinforcement Learning Using Symbolic Model Estimates

Sarath Sreedharan,Michael Katz

There has been increasing interest in using symbolic models along with reinforcement learning (RL) problems, where these coarser abstract models are used as a way to provide higher level guidance to the RL agent. However, most of these works are limited by their assumption that they have access to a symbolic approximation of the underlying problem. To address this problem, we introduce a new metho d for learning optimistic symbolic approximations of the underlying world model. We will see how these representations, coupled with fast diverse planners developed from the automated planning community, provides us with a new paradigm for optimistic exploration in sparse reward settings. We also investigate how we could speed up the learning process by generalizing learned model dynamics across similar actions with minimal human input. We will evaluate the method, by testing it on multiple benchmark domains and compare it with other RL strategies for sparse reward settings, including hierarchical RL and intrinsic reward based exploration.

****

FLIP: A Provable Defense Framework for Backdoor Mitigation in Federated Learning

Kaiyuan Zhang,Guanhong Tao,Qiuling Xu,Siyuan Cheng,Shengwei An,Yingqi Liu,Shiwei Feng,Guangyu Shen,Pin-Yu Chen,Shiqing Ma,Xiangyu Zhang

Federated Learning (FL) is a distributed learning paradigm that enables different parties to train a model together for high quality and strong privacy protection. In this scenario, individual participants may get compromised and perform backdoor attacks by poisoning the data (or gradients). Existing work on robust aggregation and certified FL robustness does not study how hardening benign clients can affect the global model (and the malicious clients). In this work, we theoretically analyze the connection among cross-entropy loss, attack success rate, and clean accuracy in this setting. Moreover, we propose a trigger reverse engineering based defense and show that our method can achieve robustness improvement with guarantee (i.e., reducing the attack success rate) without affecting benign accuracy. We conduct comprehensive experiments across different datasets and attack settings. Our results on nine competing SOTA defense methods show the empirical superiority of our method on both single-shot and continuous FL backdoor attacks. Code is available at https://github.com/KaiyuanZh/FLIP.

****

Robust attributions require rethinking robustness metrics

Sandesh Kamath,Amit Deshpande,Vineeth N. Balasubramanian

For machine learning models to be reliable and trustworthy, their decisions must be interpretable. As these models find increasing use in safety-critical applications, it is important that not just the model predictions but also their explanations (as feature attributions) be robust to small human-imperceptible input perturbations. Recent works have shown that many attribution methods are fragile and have proposed improvements in either the attribution methods or the model training. Existing works measure attributional robustness by metrics such as top-k intersection, Spearman's rank-order correlation (or Spearman's ρ) or Kendall's rank-order correlation (or Kendall's τ) to quantify the change in feature attributions under input perturbation. However, we show that these metrics are fragile. That is, under such metrics, a simple random perturbation attack can seem to be as significant as more principled attributional attacks. We instead propose Locality-sENSitive (LENS) improvements of the above metrics, namely, LENS-top-k, LENS-Spearman and LENS-Kendall, that incorporate the locality of attributions along with their rank order. Our locality-sensitive metrics provide tighter bounds on attributional robustness and do not disproportionately penalize attribution methods for reasonable local changes. We show that the robust attribution methods proposed in recent works also reflect this premise of locality, thus highlighting the need for a locality-sensitive metric for progress in the field. Our empirical results on well-known benchmark datasets using well-known models and attribution methods support our observations and conclusions in this work.

****

GradientMix: A Simple yet Effective Regularization for Large Batch Training

Jihun Yun,Jung Hyun Lee,Eunho Yang

Stochastic gradient descent (SGD) is the core tool for training deep neural networks. As modern deep learning tasks become more complex and state-of-the-art architectures grow as well, network training with SGD takes a huge amount of time; for example, training ResNet on the ImageNet dataset or BERT pre-training can take days to dozens of days. To reduce the network training time, distributed lear ning using a large batch size for SGD has been one of the main active research areas in recent years, but this approach entails a significant degradation in generalization. To address this issue, in this paper, we propose a simple yet effective regularization technique, GradientMix, for large-scale distributed learning. GradientMix can enhance the generalization in large batch regimes by giving appropriate noise through a mixup of local gradients computed at multiple devices, which is contrary to the conventions that simply average local gradients. Furthermore, GradientMix is optimizer-agnostic, hence can be applied to any popular optimization algorithm as long as the overall loss is expressed as the sum of the subgroup losses. Our extensive experiments show the effectiveness in both small and large-scale problems, and especially we consistently achieve state-of-the-art performance for various optimizers on training ResNet-50 on the ImageNet dataset with 32K batch size.

****

UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining

Hyung Won Chung,Xavier Garcia,Adam Roberts,Yi Tay,Orhan Firat,Sharan Narang,Noah Constant

Pretrained multilingual large language models have typically used heuristic temperature-based sampling to balance between different languages. However previous work has not systematically evaluated the efficacy of different pretraining language distributions across model scales. In this paper, we propose a new sampling method, UniMax, that delivers more uniform coverage of head languages while mitigating overfitting on tail languages by explicitly capping the number of repeats over each language's corpus. We perform an extensive series of ablations testing a range of sampling strategies on a suite of multilingual benchmarks, while varying model scale. We find that UniMax outperforms standard temperature-based sampling, and the benefits persist as scale increases. As part of our contribution, we release: (i) an improved and refreshed mC4 multilingual corpus consisting of 29 trillion characters across 107 languages, and (ii) a suite of pretrained umT5 model checkpoints trained with UniMax sampling.

****

Discrete State-Action Abstraction via the Successor Representation

Amnon Attali,Pedro Cisneros,Marco Morales,Nancy Amato

While the difficulty of reinforcement learning problems is typically related to the complexity of their state spaces, Abstraction proposes that solutions often lie in simpler underlying latent spaces. Prior works have focused on learning either a continuous or dense abstraction, or require a human to provide one. Information-dense representations capture features irrelevant for solving tasks, and continuous spaces can struggle to represent discrete objects. In this work we automatically learn a sparse discrete abstraction of the underlying environment. We do so using a simple end-to-end trainable model based on the successor representation and max-entropy regularization. We describe an algorithm to apply our model, named Discrete State-Action Abstraction (DSAA), which computes an action abstraction in the form of temporally extended actions, i.e., Options, to transition between discrete abstract states. Empirically, we demonstrate the effects of different exploration schemes on our resulting abstraction, and show that it is efficient for solving downstream tasks.

****

Hyper-parameter Tuning for Fair Classification without Sensitive Attribute Access

Akshaj Kumar Veldanda,Ivan Brugere,Sanghamitra Dutta,Alan Mishler,Siddharth Garg

Fair machine learning methods seek to train models that balance model performance across demographic subgroups defined over sensitive attributes like race and gender. Although sensitive attributes are typically assumed to be known during training, they may not be available in practice due to privacy and other logistical concerns. Recent work has sought to train fair models without sensitive attributes on training data. However, these methods need extensive hyper-parameter tuning to achieve good results, and hence assume that sensitive attributes are known on validation data. However, this assumption too might not be practical. Here, we propose a framework to train fair classifiers without access to sensitive attributes on either training or validation data. Instead, we generate pseudo sensitive attributes on the validation data by training a biased classifier and using the classifier’s incorrectly (correctly) labeled examples as proxies for minority (majority) groups. Since fairness metrics like demographic parity, equal opportunity and subgroup accuracy can be estimated to within a proportionality constant even with noisy sensitive attribute information, we show theoretically and empirically that these proxy labels can be used to maximize fairness under average accuracy constraints. Key to our results is a principled approach to select the hyper-parameters of the biased classifier in a completely unsupervised fashion (meaning without access to ground truth sensitive attributes) that minimizes the gap between fairness estimated using noisy versus ground-truth sensitive labels.

****

Towards Learning Implicit Symbolic Representation for Visual Reasoning

Chen Sun,Calvin Luo,Xingyi Zhou,Anurag Arnab,Cordelia Schmid

Visual reasoning tasks are designed to test a learning algorithm's capability to infer causal relationships, discover object interactions, and understand temporal dynamics, all from visual cues. It is commonly believed that to achieve compositional generalization on visual reasoning, an explicit abstraction of the visual scene must be constructed; for example, object detection can be applied to the visual input to produce representations that are then processed by a neural network or a neuro-symbolic framework. We demonstrate that a simple and general self-supervised approach is able to learn implicit symbolic representations with general-purpose neural networks, enabling the end-to-end learning of visual reasoning directly from raw visual inputs. Our proposed approach ``compresses'' each frame of a video into a small set of tokens with a transformer network. The self-supervised learning objective is to reconstruct each image based on the compressed temporal context. To minimize the reconstruction loss, the network must learn a compact representation for each image, as well as capture temporal dynamics and object permanence from temporal context. We evaluate the proposed approach on two visual reasoning benchmarks, CATER and ACRE. We observe that self-supervised pretraining is essential to achieve compositional generalization for our end-to-end trained neural network, and our proposed method achieves on par or better performance compared to recent neuro-symbolic approaches that often require additional object-level supervision.

****

GNNInterpreter: A Probabilistic Generative Model-Level Explanation for Graph Neural Networks

Xiaoqi Wang,Han Wei Shen

Recently, Graph Neural Networks (GNNs) have significantly advanced the performance of machine learning tasks on graphs. However, this technological breakthrough makes people wonder: how does a GNN make such decisions, and can we trust its prediction with high confidence? When it comes to some critical fields, such as biomedicine, where making wrong decisions can have severe consequences, it is crucial to interpret the inner working mechanisms of GNNs before applying them. In this paper, we propose a model-agnostic model-level explanation method for different GNNs that follow the message passing scheme, GNNInterpreter, to explain the high-level decision-making process of the GNN model. More specifically, GNNInterpreter learns a probabilistic generative graph distribution that produces the most discriminative graph pattern the GNN tries to detect when making a certain prediction by optimizing a novel objective function specifically designed for the model-level explanation for GNNs. Compared to existing works, GNNInterpreter is more flexible and computationally efficient in generating explanation graphs with different types of node and edge features, without introducing another blackbox or requiring manually specified domain-specific rules. In addition, the experimental studies conducted on four different datasets demonstrate that the explanation graphs generated by GNNInterpreter match the desired graph pattern if the model is ideal; otherwise, potential model pitfalls can be revealed by the explanation.

****

Rethinking Symbolic Regression: Morphology and Adaptability in the Context of Evolutionary Algorithms

Kei Sen Fong,Shelvia Wongso,Mehul Motani

Symbolic Regression (SR) is the well-studied problem of finding closed-form analytical expressions that describe the relationship between variables in a measurement dataset. In this paper, we rethink SR from two perspectives: morphology and adaptability. Morphology: Current SR algorithms typically use several man-made heuristics to influence the morphology (or structure) of the expressions in the search space. These man-made heuristics may introduce unintentional bias and data leakage, especially with the relatively few equation-recovery benchmark problems available for evaluating SR approaches. To address this, we formulate a novel minimalistic approach, based on constructing a depth-aware mathematical language model trained on terminal walks of expression trees, as a replacement to these heuristics. Adaptability: Current SR algorithms tend to select expressions based on only a single fitness function (e.g., MSE on the training set). We promote the use of an adaptability framework in evolutionary SR which uses fitness functions that alternate across generations. This leads to robust expressions that perform well on the training set and are close to the true functional form. We demonstrate this by alternating fitness functions that quantify faithfulness to values (via MSE) and empirical derivatives (via a novel theoretically justified fitness metric coined MSEDI). Proof-of-concept: We combine these ideas into a minimalistic evolutionary SR algorithm that outperforms all benchmark and state of-the-art SR algorithms in problems with unknown constants added, which we claim are more reflective of SR performance for real-world applications. Our claim is then strengthened by reproducing the superior performance on real-world regression datasets from SRBench. For researchers interested in equation-recovery problems, we also propose a set of conventions that can be used to promote fairness in comparison across SR methods and to reduce unintentional bias.

****

On Pre-training Language Model for Antibody

Danqing Wang,Fei YE,Hao Zhou

Antibodies are vital proteins offering robust protection for the human body from pathogens. The development of general protein and antibody-specific pre-trained language models both facilitate antibody prediction tasks. However, there have been limited studies that comprehensively explore the representation capability of distinct pre-trained language models on different antibody tasks. To investigate the problem, we aim to answer several key questions in this paper, such as how pre-trained language models perform in antibody tasks with different specificity and how introducing specific biological mechanisms to the pre-training process can benefit the model. Additionally, we evaluate if the learned antibody pre-trained representations can be applied to real-world antibody problems, like drug discovery and immune process understanding. Previously, no benchmark available largely hindered the study to answer these questions. To aid in our investigation, we provide an AnTibody Understanding Evaluation (ATUE) benchmark. We comprehensively evaluate the performance of protein pre-trained language models by empirical study along with conclusions and new insights. Our ATUE and code are released at https://github.com/dqwang122/EATLM.

****

Challenging Common Assumptions about Catastrophic Forgetting

Timothee LESORT,Oleksiy Ostapenko,Pau Rodriguez,Md Rifat Arefin,Diganta Misra,Laurent Charlin,Irina Rish

Standard gradient descent algorithms applied to sequences of tasks are known to induce catastrophic forgetting in deep neural networks. When trained on a new task, the model's parameters are updated in a way that degrades performance on past tasks.

This article explores continual learning (CL) on long sequences of tasks sampled from a finite environment.

\textbf{We show that in this setting, learning with stochastic gradient descent (SGD) results in knowledge retention and accumulation without specific memorizat ion mechanisms.} This is in contrast to the current notion of forgetting from the CL literature, which shows that training on new tasks with such an approach results in forgetting previous tasks, especially in class-incremental settings.

To study this phenomenon, we propose an experimental framework, \Scole{} (Scaling Continual Learning), which allows to generate arbitrarily long task sequences. Our experiments show that the previous results obtained on relatively short task sequences may not reveal certain phenomena that emerge in longer ones.

****

Learning to reason over visual objects

Shanka Subhra Mondal,Taylor Whittington Webb,Jonathan Cohen

A core component of human intelligence is the ability to identify abstract patterns inherent in complex, high-dimensional perceptual data, as exemplified by visual reasoning tasks such as Raven’s Progressive Matrices (RPM). Motivated by the goal of designing AI systems with this capacity, recent work has focused on evaluating whether neural networks can learn to solve RPM-like problems. Previous work has generally found that strong performance on these problems requires the incorporation of inductive biases that are specific to the RPM problem format, raising the question of whether such models might be more broadly useful. Here, we investigated the extent to which a general-purpose mechanism for processing visual scenes in terms of objects might help promote abstract visual reasoning. We found that a simple model, consisting only of an object-centric encoder and a transformer reasoning module, achieved state-of-the-art results on both of two challenging RPM-like benchmarks (PGM and I-RAVEN), as well as a novel benchmark with greater visual complexity (CLEVR-Matrices). These results suggest that an inductive bias for object-centric processing may be a key component of abstract visual reasoning, obviating the need for problem-specific inductive biases.

****

Imitating Graph-Based Planning with Goal-Conditioned Policies

Junsu Kim,Younggyo Seo,Sungsoo Ahn,Kyunghwan Son,Jinwoo Shin

Recently, graph-based planning algorithms have gained much attention to solve goal-conditioned reinforcement learning (RL) tasks: they provide a sequence of subgoals to reach the target-goal, and the agents learn to execute subgoal-conditioned policies. However, the sample-efficiency of such RL schemes still remains a challenge, particularly for long-horizon tasks. To address this issue, we present a simple yet effective self-imitation scheme which distills a subgoal-conditioned policy into the target-goal-conditioned policy. Our intuition here is that to reach a target-goal, an agent should pass through a subgoal, so target-goal- and subgoal- conditioned policies should be similar to each other. We also propose a novel scheme of stochastically skipping executed subgoals in a planned path, which further improves performance. Unlike prior methods that only utilize graph-based planning in an execution phase, our method transfers knowledge from a planner along with a graph into policy learning. We empirically show that our method can significantly boost the sample-efficiency of the existing goal-conditioned RL methods under various long-horizon control tasks.

****

Prefer to Classify: Improving Text Classifier via Pair-wise Preference Learning

Jaehyung Kim,Jinwoo Shin,Dongyeop Kang

The development of largely human-annotated benchmarks has driven the success of deep neural networks in various NLP tasks. These benchmarks are collected by aggregating decisions made by different annotators on the target task. Aggregating the annotated decisions via majority is still used as a common practice, despite its inevitable limitation from simple aggregation. In this paper, we establish a novel classification framework, based on task-specific human preference between a pair of samples, which provides an informative training signal to capture fine-grained and complementary task information through pair-wise comparison. Hence, it improves the existing instance-wise annotation system by enabling better task modeling from learning the relation between samples. Specifically, we propose a new multi-task learning framework, called prefer-to-classify (P2C), to effectively learn human preferences in addition to the given classification task.

We collect human preference signals in two ways: (1) extracting relative prefere nces implicitly from annotation records (for free) or (2) collecting subjective preferences explicitly from (paid) crowd workers. In various text classification tasks, we demonstrate that both extractive and subjective preferences are effective in improving the classifier with our preference learning framework. Interestingly, we found that subjective preference shows more significant improvements than extractive preference, revealing the effectiveness of explicit modeling of human preferences. Our code and preference dataset will be publicly available upon acceptance.

****

Seeing Differently, Acting Similarly: Heterogeneously Observable Imitation Learning

Xin-Qiang Cai,Yao-Xiang Ding,Zixuan Chen,Yuan Jiang,Masashi Sugiyama,Zhi-Hua Zhou

In many real-world imitation learning tasks, the demonstrator and the learner have to act under different observation spaces. This situation brings significant obstacles to existing imitation learning approaches, since most of them learn policies under homogeneous observation spaces. On the other hand, previous studies under different observation spaces have strong assumptions that these two observation spaces coexist during the entire learning process. However, in reality, the observation coexistence will be limited due to the high cost of acquiring expert observations. In this work, we study this challenging problem with limited observation coexistence under heterogeneous observations: Heterogeneously Observable Imitation Learning (HOIL). We identify two underlying issues in HOIL: the dynamics mismatch and the support mismatch, and further propose the Importance Weighting with REjection (IWRE) algorithm based on importance weighting and learning with rejection to solve HOIL problems. Experimental results show that IWRE can solve various HOIL tasks, including the challenging tasks of transforming the vision-based demonstrations to random access memory (RAM)-based policies in the Atari domain, even with limited visual observations.

****

A theoretical study of inductive biases in contrastive learning

Jeff Z. HaoChen,Tengyu Ma

Understanding self-supervised learning is important but challenging. Previous theoretical works study the role of pretraining losses, and view neural networks as general black boxes. However, the recent work of [Saunshi et al.] argues that the model architecture --- a component largely ignored by previous works --- also has significant influences on the downstream performance of self-supervised learning. In this work, we provide the first theoretical analysis of self-supervised learning that incorporates the effect of inductive biases originating from the model class. In particular, we focus on contrastive learning --- a popular self-supervised learning method that is widely used in the vision domain. We show that when the model has limited capacity, contrastive representations would recover certain special clustering structures that are compatible with the model architecture, but ignore many other clustering structures in the data distribution. As a result, our theory can capture the more realistic setting where contrastive representations have much lower dimensionality than the number of clusters in the data distribution. We instantiate our theory on several synthetic data distributions, and provide empirical evidence to support the theory.

****

Combinatorial Pure Exploration of Causal Bandits

Nuoya Xiong,Wei Chen

The combinatorial pure exploration of causal bandits is the following online learning task: given a causal graph with unknown causal inference distributions, in each round we choose a subset of variables to intervene or do no intervention, and observe the random outcomes of all random variables, with the goal that using as few rounds as possible, we can output an intervention that gives the best (or almost best) expected outcome on the reward variable Y with probability at least $1-\delta$, where δ is a given confidence level. We provide the first gap-dependent and fully adaptive pure exploration algorithms on two types of causal models --- the binary generalized linear model (BGLM) and general graphs.

For BGLM, our algorithm is the first to be designed specifically for this setting and achieves polynomial sample complexity, while all existing algorithms for general graphs have either sample complexity exponential to the graph size or some unreasonable assumptions. For general graphs, our algorithm provides a significant improvement on sample complexity, and it nearly matches the lower bound we prove. Our algorithms achieve such improvement by a novel integration of prior causal bandit algorithms and prior adaptive pure exploration algorithms, the former of which utilize the rich observational feedback in causal bandits but are not adaptive to reward gaps, while the latter of which have the issue in reverse.
****

How to fine-tune vision models with SGD
Ananya Kumar,Ruoqi Shen,Sebastien Bubeck,Suriya Gunasekar
SGD (with momentum) and AdamW are the two most commonly used optimizers for fine-tuning large neural networks in computer vision. When the two methods perform the same, SGD is preferable because it uses less memory and is more efficient than AdamW. However, when evaluating on downstream tasks that differ significantly from pretraining, we find that across five popular benchmarks SGD fine-tuning gets substantially lower accuracies than AdamW on many modern vision models such as Vision Transformers and ConvNeXts---especially out-of-distribution (OOD). We find that such large gaps arise in instances where the fine-tuning gradients in the first (``embedding'') layer are much larger than the rest of the model. Our analysis suggests an easy fix: if we simply freeze the embedding layer (0.7\% of the parameters), SGD performs competitively with AdamW while using less memory across a suite of benchmarks. Our insights lead to state-of-the-art accuracies on popular distribution shift benchmarks: WILDS-FMoW, WILDS-Camelyon, BREEDS-Living-17, Waterbirds, and DomainNet.
****

Computational Language Acquisition with Theory of Mind
Andy Liu,Hao Zhu,Emmy Liu,Yonatan Bisk,Graham Neubig
Unlike current state-of-the-art language models, young children actively acquire language through interactions with their surrounding environment and caretakers. One mechanism that has been argued to be critical to language learning is the ability to infer the mental states of other agents in social environments, coined Theory of Mind (ToM) by Premack & Woodruff (1978). Drawing inspiration from the modern operationalized versions of ToM implemented in Rabinowitz et al. (2018) and Zhu et al. (2021), we build language-learning agents equipped with ToM, and measure its effects on the learning process. We model ToM by giving the speaker agent an internal listener model that is trained alongside the speaker and used to rerank potential utterances. We experiment with varying task difficulty, hypothesizing that models will acquire more complex language to adapt to stronger environmental pressures. We find that training speakers with a highly weighted ToM listener component leads to performance gains in our image referential game setting. We also find some evidence that increasing task difficulty in the training process results in more fluent and precise utterances in evaluation. This suggests the potential utility of further incorporating ToM, as well as other insights from child language acquisition, into computational models of language acquisition.
****

MiDAS: Multi-integrated Domain Adaptive Supervision for Fake News Detection
Abhijit Suprem,Calton Pu
COVID-19 related misinformation and fake news, coined an 'infodemic', has dramatically increased over the past few years. This misinformation exhibits concept drift, where the distribution of fake news changes over time, reducing effectiveness of previously trained models for fake news detection. Given a set of fake news models trained on multiple domains, we propose an adaptive decision module to select the best-fit model for a new sample. We propose MiDAS, a multi-domain adaptative approach for fake news detection that ranks relevancy of existing models to new samples. MiDAS contains 2 components: a doman-invariant encoder, and an adaptive model selector. MiDAS integrates multiple pre-trained and fine-tuned models with their training data to create a domain-invariant representation. Then

, MiDAS uses local Lipschitz smoothness of the invariant embedding space to estimate each model's relevance to a new sample. Higher ranked models provide predictions, and lower ranked models abstain. We evaluate MiDAS on generalization to drifted data with 9 fake news datasets, each obtained from different domains and modalities. MiDAS achieves new state-of-the-art performance on multi-domain adaptation for out-of-distribution fake news classification.

****

Walking the Tightrope: An Investigation of the Convolutional Autoencoder Bottleneck

Ilja Manakov,Markus Rohm,Volker Tresp

In this paper, we present an in-depth investigation of the convolutional autoencoder (CAE) bottleneck.

Autoencoders (AE), and especially their convolutional variants, play a vital role in the current deep learning toolbox.

Researchers and practitioners employ CAEs for various tasks, ranging from outlier detection and compression to transfer and representation learning.

Despite their widespread adoption, we have limited insight into how the bottleneck shape impacts the CAE's emergent properties.

We demonstrate that increased bottleneck area (i.e., height $\times$ width) drastically improves generalization in terms of reconstruction error while also speeding up training.

The number of channels in the bottleneck, on the other hand, is of secondary importance.

Furthermore, we show empirically that CAEs do not learn to copy their input, even when all layers have the same number of neurons as there are pixels in the input (i.e. there is no bottleneck).

Besides raising important questions for further research, our findings are directly applicable to two of the most common use-cases for CAEs:

In image compression, it is advantageous to increase the feature map size in the bottleneck as this greatly improves reconstruction quality.

For reconstruction-based outlier detection, we recommend decreasing the feature map size so that out-of-distribution samples will yield a higher reconstruction error.

****

A Closer Look at Model Adaptation using Feature Distortion and Simplicity Bias

Puja Trivedi,Danai Koutra,Jayaraman J. Thiagarajan

Advances in the expressivity of pretrained models have increased interest in the design of adaptation protocols which enable safe and effective transfer learning. Going beyond conventional linear probing (LP) and fine tuning (FT) strategies, protocols that can effectively control feature distortion, i.e., the failure to update features orthogonal to the in-distribution, have been found to achieve improved out-of-distribution generalization (OOD). In order to limit this distortion, the LP+FT protocol, which first learns a linear probe and then uses this initialization for subsequent FT, was proposed. However, in this paper, we find when adaptation protocols (LP, FT, LP+FT) are also evaluated on a variety of safety objectives (e.g., calibration, robustness, etc.), a complementary perspective to feature distortion is helpful to explain protocol behavior. To this end, we study the susceptibility of protocols to simplicity bias (SB), i.e. the well-known propensity of deep neural networks to rely upon simple features, as SB has recently been shown to underlie several problems in robust generalization. Using a synthetic dataset, we demonstrate the susceptibility of existing protocols to SB. Given the strong effectiveness of LP+FT, we then propose modified linear probes that help mitigate SB, and lead to better initializations for subsequent FT. We verify the effectiveness of the proposed LP+FT variants for decreasing SB in a controlled setting, and their ability to improve OOD generalization and safety on three adaptation datasets.

****

Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in Out-of-Distribution Generalization

Yongqiang Chen,Kaiwen Zhou,Yatao Bian,Binghui Xie,Bingzhe Wu,Yonggang Zhang,MA K

AILI,Han Yang,Peilin Zhao,Bo Han,James Cheng

Recently, there has been a growing surge of interest in enabling machine learning systems to generalize well to Out-of-Distribution (OOD) data. Most efforts are devoted to advancing optimization objectives that regularize models to capture the underlying invariance; however, there often are compromises in the optimization process of these OOD objectives: i) Many OOD objectives have to be relaxed as penalty terms of Empirical Risk Minimization (ERM) for the ease of optimization, while the relaxed forms can weaken the robustness of the original objective; ii) The penalty terms also require careful tuning of the penalty weights due to the intrinsic conflicts between ERM and OOD objectives. Consequently, these compromises could easily lead to suboptimal performance of either the ERM or OOD objective. To address these issues, we introduce a multi-objective optimization (MOO) perspective to understand the OOD optimization process, and propose a new optimization scheme called PAreto Invariant Risk Minimization (PAIR). PAIR improves the robustness of OOD objectives by cooperatively optimizing with other OOD objectives, thereby bridging the gaps caused by the relaxations. Then PAIR approaches a Pareto optimal solution that trades off the ERM and OOD objectives properly. Extensive experiments on challenging benchmarks, WILDS, show that PAIR alleviates the compromises and yields top OOD performances.

****

Understanding and Adopting Rational Behavior by Bellman Score Estimation

Kuno Kim,Stefano Ermon

We are interested in solving a class of problems that seek to understand and adopt rational behavior from demonstrations. We may broadly classify these problems into four categories of reward identification, counterfactual analysis, behavior imitation, and behavior transfer. In this work, we make a key observation that knowing how changes in the underlying rewards affect the optimal behavior allows one to solve a variety of aforementioned problems. To a local approximation, this quantity is precisely captured by what we term the Bellman score, i.e gradient of log probabilities of the optimal policy with respect to the reward. We introduce the Bellman score operator which provably converges to the gradient of the infinite-horizon optimal Q-values with respect to the reward which can then be used to directly estimate the score. Guided by our theory, we derive a practical score-learning algorithm which can be used for score estimation in high-dimensional state-actions spaces. We show that score-learning can be used to reliably identify rewards, perform counterfactual predictions, achieve state-of-the-art behavior imitation, and transfer policies across environments.

****

L2B: Learning to Bootstrap for Combating Label Noise

Yuyin Zhou,Xianhang Li,Fengze Liu,Xuxi Chen,Lequan Yu,Cihang Xie,Matthew P. Lungren,Lei Xing

Deep neural networks are powerful tools for representation learning, but can easily overfit to noisy labels which are prevalent in many real-world scenarios. Generally, noisy supervision could stem from variation among labelers, label corruption by adversaries, etc. To combat such label noises, one popular line of approach is to apply customized weights to the training instances, so that the corrupted examples contribute less to the model learning. However, such learning mechanisms potentially erase important information about the data distribution and therefore yield suboptimal results. To leverage useful information from the corrupted instances, an alternative is the bootstrapping loss, which reconstructs new training targets on-the-fly by reweighting the real labels and the network's own predictions (i.e., pseudo labels).

In this paper, we propose a more generic learnable loss objective which enables a joint reweighting of instances and labels at once. Specifically, our method dynamically adjusts the $\textit{per-sample importance weight}$ between the real observed labels and pseudo-labels, where the weights are efficiently determined in a meta process. Compared to the previous instance reweighting methods, our approach concurrently conducts implicit relabeling, and thereby yields substantial improvements with almost no extra cost. Extensive experimental results demonstrated the strengths of our approach over existing methods on multiple natural and medical image benchmark datasets, including CIFAR-10, CIFAR-100, ISIC2019 and Clothing 1M. Code will be made publicly available.

****

What Makes Convolutional Models Great on Long Sequence Modeling?
Yuhong Li,Tianle Cai,Yi Zhang,Deming Chen,Debadeepta Dey

Convolutional models have been widely used in multiple domains. However, most existing models only use local convolution, making the model unable to handle long-range dependencies efficiently. Attention overcomes this problem by aggregating global information based on the pair-wise attention score but also makes the computational complexity quadratic to the sequence length. Recently, Gu et al. proposed a model called S4 inspired by the state space model. S4 can be efficiently implemented as a global convolutional model whose kernel size equals the input sequence length. With Fast Fourier Transform, S4 can model much longer sequences than Transformers and achieve significant gains over SoTA on several long-range tasks. Despite its empirical success, S4 is involved. It requires sophisticated parameterization and initialization schemes that combine the wisdom from several prior works. As a result, S4 is less intuitive and hard to use for researchers with limited prior knowledge. Here we aim to demystify S4 and extract basic principles that contribute to the success of S4 as a global convolutional model. We focus on the structure of the convolution kernel and identify two critical but intuitive principles enjoyed by S4 that are sufficient to make up an effective global convolutional model: 1) The parameterization of the convolutional kernel needs to be efficient in the sense that the number of parameters should scale sub-linearly with sequence length. 2) The kernel needs to satisfy a decaying structure that the weights for convolving with closer neighbors are larger than the more distant ones. Based on the two principles, we propose a simple yet effective convolutional model called Structured Global Convolution (SGConv). SGConv exhibits strong empirical performance over several tasks: 1) With faster speed, SGConv surpasses the previous SoTA on Long Range Arena and Speech Command datasets. 2) When plugging SGConv into standard language and vision models, it shows the potential to improve both efficiency and performance.

****

Progressive Mixup Augmented Teacher-Student Learning for Unsupervised Domain Adaptation
Aotian Zheng,Jie Mei,Farron Wallace,Craig Rose,Rania Hussein,Jenq-Neng Hwang

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain, mostly through learning a domain invariant feature representation. Currently, the best performing UDA methods use category level domain alignment to capture fine-grained information, resulting in significantly improved performance over global alignment. While successful, category level UDA methods suffer from the unreliable pseudo-labels for target data. Additionally, most UDA methods directly adapt from source to target domain without regard for the large domain discrepancy. In this paper, we propose an UDA approach with teacher-student learning where the teacher network is used to provide more reliable target pseudo-labels for the student network to train with. Furthermore, we use a progressive mixup augmentation strategy which generates intermediate samples that become increasingly target-dominant as training progresses. Aligning the source and intermediate domains allows the model to gradually transfer fine-grained domain knowledge from the source to the target domain while minimizing the negative impact of noisy target pseudo-labels. This progressive mixup augmented teacher-student (PMATS) training strategy along with class subset sampling and clustering based pseudo-label refinement achieves state-of-the-art performance on two public UDA benchmark datasets: Office-31, and Office-Home.

****

M^{3}SAT: A Sparsely Activated Transformer for Efficient Multi-Task Learning from Multiple Modalities
Jie Peng,Tianlong Chen,Ruida Zhou,Jianmin Ji,Yanyong Zhang,Zhangyang Wang

Multi-modal multi-task learning (M^{2}TL) aims to discover the implicit correspondences among heterogeneous modalities and tasks, which is common in real-world applications like autonomous driving and robotics control. Current single-model solutions for M^{2}TL usually fall short in several aspects. The shared backbone between the modalities is prone to overfitting the simpler modality, while jointly optimizing the tasks suffers from unstable training due to the gradient conflicts across tasks. On the other hand, designing a separate model for each task and modality can avoid the above problems but leads to prohibitively expensive computation and memory consumption, rendering this approach unrealistic.

In this work, we propose M^{3}SAT, a sparsely activated transformer for efficient M^{2}TL. The proposed framework tailors the mixture-of-experts (MoEs) into both the self-attention and the feed-forward networks (FFN) of a transformer backbone. It adopts the routing policy to assign attention-heads and FFN experts during training, which effectively disentangles the parameter space to prevent training conflicts among diverse modalities and tasks. Meanwhile, disentangled parameter space also restrains the problem of simple modal prone to overfitting. Sparsely activating the transformer also enables efficient computation for each input sample. Through comprehensive evaluation, we demonstrate the effectiveness of our M^{3}SAT: a remarkable performance margin (\textit{e.g.}, $\ge 1.37\%$) is achieved over the dense models with the same computation cost. More importantly, M^{3}SAT can achieve the above performance improvements with a fraction of the computation cost -- our computation is only $1.38\% \sim 53.51\%$ of that of the SOTA methods. Our code will be released upon acceptance.
****
Editing models with task arithmetic
Gabriel Ilharco,Marco Tulio Ribeiro,Mitchell Wortsman,Ludwig Schmidt,Hannaneh Hajishirzi,Ali Farhadi
Changing how pre-trained models behave---e.g., improving their performance on a downstream task or mitigating biases learned during pre-training---is a common practice when developing machine learning systems. In this work, we propose a new paradigm for steering the behavior of neural networks, centered around task vectors. A task vector specifies a direction in the weight space of a pre-trained model, such that movement in that direction improves performance on the task. We build task vectors by subtracting the weights of a pre-trained model from the weights of the same model after fine-tuning on a task. We show that these task vectors can be modified and combined together through arithmetic operations such as negation and addition, and the behavior of the resulting model is steered accordingly. Moreover, task vectors can be added together to improve performance on multiple tasks at once. Finally, when tasks are linked by an analogy relationship of the form ``A is to B as C is to D", combining task vectors from three of the tasks can improve performance on the fourth, even when no data from the fourth task is used for training.
****
Neural Systematic Binder
Gautam Singh,Yeongbin Kim,Sungjin Ahn
The key to high-level cognition is believed to be the ability to systematically manipulate and compose knowledge pieces. While token-like structured knowledge representations are naturally provided in text, it is elusive how to obtain them for unstructured modalities such as scene images. In this paper, we propose a neural mechanism called Neural Systematic Binder or SysBinder for constructing a novel structured representation called Block-Slot Representation. In Block-Slot Representation, object-centric representations known as slots are constructed by composing a set of independent factor representations called blocks, to facilitate systematic generalization. SysBinder obtains this structure in an unsupervised way by alternatingly applying two different binding principles: spatial binding for spatial modularity across the full scene and factor binding for factor modularity within an object. SysBinder is a simple, deterministic, and general-purpose layer that can be applied as a drop-in module in any arbitrary neural network and on any modality. In experiments, we find that SysBinder provides significantly better factor disentanglement within the slots than the conventional object-centric methods, including, for the first time, in visually complex scene imag

es such as CLEVR-Tex. Furthermore, we demonstrate factor-level systematicity in controlled scene generation by decoding unseen factor combinations.

****

Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis

Weixi Feng,Xuehai He,Tsu-Jui Fu,Varun Jampani,Arjun Reddy Akula,Pradyumna Narayana,Sugato Basu,Xin Eric Wang,William Yang Wang

Large-scale diffusion models have achieved state-of-the-art results on text-to-image synthesis (T2I) tasks. Despite their ability to generate high-quality yet creative images, we observe that attribution-binding and compositional capabilities are still considered major challenging issues, especially when involving multiple objects. Attribute-binding requires the model to associate objects with the correct attribute descriptions, and compositional skills require the model to combine and generate multiple concepts into a single image. In this work, we improve these two aspects of T2I models to achieve more accurate image compositions. To do this, we incorporate linguistic structures with the diffusion guidance process based on the controllable properties of manipulating cross-attention layers in diffusion-based T2I models. We observe that keys and values in cross-attention layers have strong semantic meanings associated with object layouts and content. Therefore, by manipulating the cross-attention representations based on linguistic insights, we can better preserve the compositional semantics in the generated image. Built upon Stable Diffusion, a SOTA T2I model, our structured cross-attention design is efficient that requires no additional training samples. We achieve better compositional skills in qualitative and quantitative results, leading to a significant 5-8\% advantage in head-to-head user comparison studies. Lastly, we conduct an in-depth analysis to reveal potential causes of incorrect image compositions and justify the properties of cross-attention layers in the generation process.

****

Atomized Deep Learning Models

Yi-Lin Tuan,Zih-Yun Chiu,William Yang Wang

Deep learning models often tackle the intra-sample structure, such as the order of words in a sentence and pixels in an image, but have not pay much attention to the inter-sample relationship. In this paper, we show that explicitly modeling the inter-sample structure to be more discretized can potentially help model's expressivity. We propose a novel method, Atom Modeling, that can discretize a continuous latent space by drawing an analogy between a data point and an {\it atom}, which is naturally spaced away from other atoms with distances depending on their intra structures. Specifically, we model each data point as an atom composed of electrons, protons, and neutrons and minimize the potential energy caused by the interatomic force among data points. Through experiments with qualitative analysis in our proposed Atom Modeling on synthetic and real datasets, we find that Atom Modeling can improve the performance by maintaining the inter-sample relation and can capture an interpretable intra-sample relation by mapping each component in a data point to electron/proton/neutron.

****

Topology Matters in Fair Graph Learning: a Theoretical Pilot Study

Zhimeng Jiang,Xiaotian Han,Chao Fan,Zirui Liu,Xiao Huang,Na Zou,Ali Mostafavi,Xia Hu

Recent advances in fair graph learning observe that graph neural networks (GNNs) further amplify prediction bias compared with multilayer perception (MLP), while the reason behind this is unknown. In this paper, we conduct a theoretical analysis of the bias amplification mechanism in GNNs. This is a challenging task since GNNs are difficult to be interpreted, and real-world networks are complex. To bridge the gap, we theoretically and experimentally demonstrate that aggregation operation in representative GNNs accumulates bias in node representation due to topology bias induced by graph topology. We provide a sufficient condition identifying the statistical information of graph data, so that graph aggregation enhances prediction bias in GNNs.

Motivated by this data-centric finding, we propose a fair graph refinement algo rithm, named \textit{FairGR}, to rewire graph topology to reduce sensitive homophily coefficient while preserving useful graph topology. Experiments on node classification tasks demonstrate that \textit{FairGR} can mitigate the prediction bias with comparable performance on three real-world datasets. Additionally, \textit{FairGR} is compatible with many state-of-the-art methods, such as adding regularization, adversarial debiasing, and Fair mixup via refining graph topology. Therefore, \textit{FairGR} is a plug-in fairness method and can be adapted to improve existing fair graph learning strategies.

****

Context-Aware Image Completion

Jinoh Cho,Minguk Kang,Vibhav Vineet,Jaesik Park

Image completion is a task that aims to fill in the missing region of a masked image with plausible contents. However, existing image completion methods tend to fill in the missing region with the surrounding texture instead of hallucinating a visual instance that is suitable in accordance with the context of the scene. In this work, we propose a novel image completion model, dubbed Refill, that hallucinates the missing instance that harmonizes well with - and thus preserves - the original context. Refill first adopts a transformer architecture that considers the types, locations of the visible instances, and the location of the missing region. Then, Refill completes the missing foreground and background semantic segmentation masks within the missing region, providing pixel-level semantic and structural guidance to generate missing contents with seamless boundaries. Finally, we condition the image synthesis blocks by using the completed segmentation mask to generate photo-realistic contents to fill out the missing region. Experimental results show the superiority of Refill over state-of-the-art image completion approaches on various natural images.

****

Can Agents Run Relay Race with Strangers? Generalization of RL to Out-of-Distribution Trajectories

Li-Cheng Lan,Huan Zhang,Cho-Jui Hsieh

In this paper, we evaluate and improve the generalization performance for reinforcement learning (RL) agents on the set of ``controllable'' states, where good policies exist on these states to achieve the goal. An RL agent that generally masters a task should reach its goal starting from any controllable state of the environment instead of memorizing a small set of trajectories. To practically evaluate this type of generalization, we propose relay evaluation, which starts the test agent from the middle of other independently well-trained stranger agents' trajectories. With extensive experimental evaluation, we show the prevalence of generalization failure on controllable states from stranger agents. For example, in the Humanoid environment, we observed that a well-trained Proximal Policy Optimization (PPO) agent, with only 3.9\% failure rate during regular testing, failed on 81.6\% of the states generated by well-trained stranger PPO agents. To improve "relay generalization," we propose a novel method called Self-Trajectory Augmentation (STA), which will reset the environment to the agent's old states according to the Q function during training. After applying STA to the Soft Actor Critic's (SAC) training procedure, we reduced the failure rate of SAC under relay-evaluation by more than three times in most settings without impacting agent performance and increasing the needed number of environment interactions. Our code is available at https://github.com/lan-lc/STA.

****

DYNAMIC BATCH NORM STATISTICS UPDATE FOR NATURAL ROBUSTNESS

Shahbaz Rezaei,Mohammad Sadegh Norouzzadeh

DNNs trained on natural clean samples have been shown to perform poorly on corrupted samples,
such as noisy or blurry images. Various data augmentation methods have been recently proposed
to improve DNN’s robustness against common corruptions. Despite their success, they require
computationally expensive training and cannot be applied to off-the-shelf trained models. Recently,

updating only BatchNorm(BN) statistics of a model on a single corruption has been shown to improve its accuracy on that corruption significantly. However, adopting the idea at inference time when the type of corruption changes decreases the effectiveness of this method. In this paper, we harness the Fourier domain to detect the corruption type, a challenging task in the image domain. We propose a unified framework consisting of a corruption-detection model and BN statistics update that can improve the corruption accuracy of any off-the-shelf trained model. We benchmark our framework on different models and datasets. Our results demonstrate about 8% and 4% accuracy improvement on CIFAR10-C and ImageNet-C, respectively. Furthermore, our framework can further improve the accuracy of state-of-the-art robust models, such as AugMix and DeepAug.

****

SKTformer: A Skeleton Transformer for Long Sequence Data
xue wang,Tian Zhou,Jianqing Zhu,Jialin Liu,Kun Yuan,Tao Yao,Wotao Yin,Rong Jin,HanQin Cai

Transformers have become a preferred tool for modeling sequential data. Many studies of using Transformers for long sequence modeling focus on reducing computational complexity. They usually exploit the low-rank structure of data and approximate a long sequence by a sub-sequence. One challenge with such approaches is how to make an appropriate tradeoff between information preserving and noise reduction: the longer the sub-sequence used to approximate the long sequence, the better the information is preserved but at a price of introducing more noise into the model and of course more computational costs. We propose skeleton transformer, SKTformer for short, an efficient transformer architecture that effectively addresses the tradeoff. It introduces two mechanisms to effectively reduce the impact of noise while still keeping the computation linear to the sequence length: a smoothing block to mix information over long sequences and a matrix sketch method that simultaneously selects columns and rows from the input matrix. We verify the effectiveness of SKTformer both theoretically and empirically. Extensive studies over both Long Range Arena (LRA) datasets and six time-series forecasting show that SKTformer significantly outperforms both villain Transformer and other state-of-the-art variants of Transformer. Code is available at
https://anonymous.4open.science/r/SKTFormer-B33B/

****

CktGNN: Circuit Graph Neural Network for Electronic Design Automation
Zehao Dong,Weidong Cao,Muhan Zhang,Dacheng Tao,Yixin Chen,Xuan Zhang

The electronic design automation of analog circuits has been a longstanding challenge in the integrated circuit field due to the huge design space and complex design trade-offs among circuit specifications. In the past decades, intensive research efforts have only been paid to automate the transistor sizing with a given circuit topology. By recognizing the graph nature of circuits, this paper presents a Circuit Graph Neural Network (CktGNN) that simultaneously automates the circuit topology generation and device sizing based on the encoder-dependent optimization subroutines. Particularly, CktGNN encodes circuit graphs using a two-level GNN framework (of nested GNN) where circuits are represented as combinations of subgraphs in a known subgraph basis. In this way, it significantly improves efficiency by reducing the number of subgraphs to perform message passing.

Nonetheless, another critical roadblock to advancing learning-assisted circuit design automation is a lack of public benchmarks to perform canonical assessment and reproducible research. To tackle the challenge, we introduce Open Circuit Benchmark (OCB), an open-sourced dataset that contains 10K distinct operational amplifiers with carefully-extracted circuit specifications from physical implementations. OCB also equips with communicative circuit generation and evaluation c apabilities such that it can be used to generalize the applicability of CktGNN to design various analog circuits by efficiently producing corresponding datasets. Experiments on OCB show the extraordinary advantages of CktGNN through representation-based optimization frameworks over other recent powerful GNN baselines and manual design from human experts. Our work paves the way toward a learning-based open-sourced design automation flow for analog circuits.

****

Substructure-Atom Cross Attention for Molecular Representation Learning
Jiye Kim,Seungbeom Lee,Dongwoo Kim,Sungsoo Ahn,Jaesik Park

Designing a neural network architecture for molecular representation is crucial for AI-driven drug discovery and molecule design. In this work, we propose a new framework for molecular representation learning. Our contribution is threefold: (a) demonstrating the usefulness of incorporating substructures to node-wise features from molecules, (b) designing two branch networks consisting of a transformer and a graph neural network so that the networks fused with asymmetric attention, and (c) not requiring heuristic features and computationally-expensive information from molecules. Using 1.8 million molecules collected from ChEMBL and PubChem database, we pretrain our network to learn a general representation of molecules with minimal supervision. The experimental results show that our pretrained network achieves competitive performance on 11 downstream tasks for molecular property prediction.

****

Differentially Private Algorithms for Smooth Nonconvex ERM
Changyu Gao,Stephen Wright

We develop simple differentially private optimization algorithms that move along directions of (expected) descent to find approximate second-order necessary solution for non-convex ERM problems. We use line search, mini-batching, and a two-phase strategy to improve the speed and practicality of the algorithm. Numerical experiments demonstrate the effectiveness of these approaches.

****

Untangling Effect and Side Effect: Consistent Causal Inference in Non-Targeted Trials
Georgios Mavroudeas,Malik Magdon-Ismail,Kristin Bennett,Jason Kuruzovich

A treatment is usually appropriate for some group (the ``sick" group) on whom it has an effect, but it can also have a side-effect when given to subjects from another group (the ``healthy" group). In a non-targeted trial both sick and healthy subjects may be treated, producing
heterogeneous effects within the treated group. Inferring the correct treatment effect on the sick population is then
difficult, because the effect and side-effect are tangled. We propose an efficient nonparametric approach to untangling the effect and side-effect, called PCM (pre-cluster and merge). We prove its asymptotic consistency in a general setting and
show, on synthetic data,
more than a 10x improvement in accuracy over existing state-of-the-art.

****

STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables
Jaehyun Nam,Jihoon Tack,Kyungmin Lee,Hankook Lee,Jinwoo Shin

Learning with few labeled tabular samples is often an essential requirement for industrial machine learning applications as varieties of tabular data suffer from high annotation costs or have difficulties in collecting new samples for novel tasks. Despite the utter importance, such a problem is quite under-explored in the field of tabular learning, and existing few-shot learning schemes from other domains are not straightforward to apply, mainly due to the heterogeneous characteristics of tabular data. In this paper, we propose a simple yet effective framework for few-shot semi-supervised tabular learning, coined Self-generated Tasks from UNlabeled Tables (STUNT). Our key idea is to self-generate diverse few-shot tasks by treating randomly chosen columns as a target label. We then employ a meta-learning scheme to learn generalizable knowledge with the constructed task

s. Moreover, we introduce an unsupervised validation scheme for hyperparameter search (and early stopping) by generating a pseudo-validation set using STUNT from unlabeled data. Our experimental results demonstrate that our simple framework brings significant performance gain under various tabular few-shot learning benchmarks, compared to prior semi- and self-supervised baselines. Code is available at https://github.com/jaehyun513/STUNT.

****

The Role of Pre-training Data in Transfer Learning

Rahim Entezari,Mitchell Wortsman,Olga Saukh,M. Moein Shariatnia,Hanie Sedghi,Ludwig Schmidt

The transfer learning paradigm of model pre-training and subsequent fine-tuning produces high accuracy models. However, a question remains: what data and method should be used for pre-training? We study the effect of the pre-training distribution on transfer learning in the context of image classification. Through controlled experiments, we find that the pre-training dataset is initially important for low-shot transfer. However, the differences between distributions is diminished as more data is made available for fine-tuning. Still, fine-tuning outperforms training from scratch. We also investigate dataset size and observe that larger pre-training datasets lead to better accuracy, however, the absolute accuracy difference is largest in the few-shot regime. Beyond data, we study the effect of the pre-training method, language-image contrastive vs. image-image contrastive, finding that the latter usually leads to better transfer accuracy

****

Compressed Predictive Information Coding

Rui Meng,Tianyi Luo,Kristofer Bouchard

Unsupervised learning plays an important role in many fields, such as machine learning, data compression, and neuroscience. Compared to static data, methods for extracting low-dimensional structure for dynamic data are lagging. We developed a novel information-theoretic framework, Compressed Predictive Information Coding (CPIC), to extract predictive latent representations from dynamic data. Predictive information quantifies the ability to predict the future of a time series from its past. CPIC selectively projects the past (input) into a low dimensional space that is predictive about the compressed data projected from the future (output). The key insight of our framework is to learn representations by balancing the minimization of compression complexity with maximization of the predictive information in the latent space. We derive tractable variational bounds of the CPIC loss by leveraging bounds on mutual information. The CPIC loss induces the latent space to capture information that is maximally predictive of the future of the data from the past. We demonstrate that introducing stochasticity in the encoder and maximizing the predictive information in latent space contributes to learning more robust latent representations. Furthermore, our variational approaches perform better in mutual information estimation compared with estimates under the Gaussian assumption commonly used. We show numerically in synthetic data that CPIC can recover dynamical systems embedded in noisy observation data with low signal-to-noise ratio. Finally, we demonstrate that CPIC extracts features more predictive of forecasting exogenous variables as well as auto-forecasting in various real datasets compared with other state-of-the-art representation learning models. Together, these results indicate that CPIC will be broadly useful for extracting low-dimensional dynamic structure from high-dimensional, noisy time-series data.

****

Importance of Class Selectivity in Early Epochs of Training

Omkar Ranadive,Nikhil Thakurdesai,Matthew L Leavitt,Stephane Deny

Deep networks trained for classification exhibit class-selective neurons in intermediate layers. Intriguingly, recent studies have shown that class-selective neurons are not strictly necessary for network function. But if class-selective neurons are not necessary, why do they exist? We attempt to answer this question in a series of experiments on ResNet-50 trained on ImageNet. We begin by showing that class-selective neurons emerge in the first few epochs of training before receding rapidly. Single-neuron ablation experiments show that class-selective ne

urons are important for network function during this early phase of training. The network is close to a linear regime during this early training phase, which may explain the emergence of these class-selective neurons in intermediate layers. Finally, by regularizing against class selectivity at different points in training, we show that the emergence of these class-selective neurons during the first few epochs of training is essential to the successful training of the network. Altogether, our results indicate that class-selective neurons in intermediate layers are vestigial remains of early epochs of training, during which they appear as quasi-linear shortcut solutions to the classification task which are essential to the successful training of the network.
****

Mechanistic Mode Connectivity
Ekdeep Singh Lubana,Eric J Bigelow,Robert Dick,David Krueger,Hidenori Tanaka
With the rise of pretrained models, fine-tuning has become of central importance in deep learning. However, unlike retraining from scratch, fine-tuning can fail to qualitatively change the behavior of a pre-trained network. For instance, we find in practice that naive fine-tuning does not eliminate a model’s sensitivity to spurious features. To understand and address this limitation, we study the geometry of neural network loss landscapes through the lens of mode-connectivity. Our work addresses two questions about mode-connectivity: 1) Are models trained on different data distributions mode-connected? 2) Can we fine tune a pre-trained model to switch modes? We define a notion of mechanistic mode-connectivity, and find that only models that already share the same invariances (which we call “mechanistically similar”) are mechanistically mode-connected. We hypothesize this property explains inability of naive fine-tuning methods to induce invariance to spurious features. Based on our analysis, we propose and validate a method of “mechanistic fine-tuning” called connectivity-based fine-tuning (CBFT)
****

WebBrain: Learning to Generate Factually Correct Articles for Queries by Grounding on Large Web Corpus
Hongjin Qian,Yutao Zhu,Zhicheng Dou,Haoqi Gu,Xinyu Zhang,Zheng Liu,Ruofei Lai,Zhao Cao,Jian-Yun Nie,Ji-Rong Wen
In this paper, we introduce a new NLP task – generating short factual articles for queries by mining supporting evidence from the Web. In this task, called WebBrain, the ultimate goal is to generate a fluent, informative, and factually-correct short article (e.g., Wiki article) for a factual query unseen in Wikipedia. To enable experiments on WebBrain, we construct a large-scale dataset WebBrain-Raw by extracting English Wikipedia articles and their crawlable Wiki references. WebBrain-Raw is ten times larger than the previous biggest peer dataset, which can greatly benefit the research community. Besides, we empirically analyze the performances of the current state-of-the-art NLP techniques on WebBrain and introduce a new framework ReGen, which enhances the generation factualness by improved evidence retrieval and task-specific pre-training for generation. Experiment results show that ReGen outperforms all baselines in both automatic and human evaluations.
****

HloEnv: A Graph Rewrite Environment for Deep Learning Compiler Optimization Research
Chin Yang Oh,Kunhao Zheng,Bingyi Kang,Xinyi Wan,Zhongwen Xu,Shuicheng YAN,Min Lin,Yangzihao Wang
We introduce HloEnv, an environment based on Accelerated Linear Algebra (XLA) for deep learning (DL) compiler optimization research. HloEnv transforms all graph rewrites into a common representation, providing a flexible interface to control and modify existing graph optimization passes. In this representation, an XLA pass is converted into a set of sequential rewrite decisions, which control when and if the rewrites are applied. Along with HloEnv, we present a dataset with broad coverage of computation graphs drawn from modern real-world machine learning models. We select two XLA passes with the largest impact on the runtime of the compiled program, and explore the potential for further improvement over XLA in this decision space. We show that using simple heuristics for decision-making c

an achieve on-par or better performance than XLA. Using search algorithms further boosts performance. We intend for HloEnv and our dataset to be an open-source, community-driven effort that helps spur advances in DL compiler optimization research.

****

Deep Latent State Space Models for Time-Series Generation

Linqi Zhou,Michael Poli,Winnie Xu,Stefano Massaroli,Stefano Ermon

Methods based on ordinary differential equations (ODEs) are widely used to build generative models of time-series. In addition to high computational overhead due to explicitly computing hidden states recurrence, existing ODE-based models fall short in learning sequence data with sharp transitions - common in many real-world systems - due to numerical challenges during optimization. In this work, we propose LS4, a generative model for sequences with latent variables evolving according to a state space ODE to increase modeling capacity. Inspired by recent deep state space models (S4), we achieve speedups by leveraging a convolutional representation of LS4 which bypasses the explicit evaluation of hidden states. We show that LS4 significantly outperforms previous continuous-time generative models in terms of marginal distribution, classification, and prediction scores on real-world datasets in the Monash Forecasting Repository, and is capable of modeling highly stochastic data with sharp temporal transitions. LS4 sets state-of-the-art for continuous-time latent generative models, with significant improvement of mean squared error and tighter variational lower bounds on irregularly-sampled datasets, while also being x100 faster than other baselines on long sequences.

****

Specformer: Spectral Graph Neural Networks Meet Transformers

Deyu Bo,Chuan Shi,Lele Wang,Renjie Liao

Spectral graph neural networks (GNNs) learn graph representations via spectral-domain graph convolutions. However, most existing spectral graph filters are scalar-to-scalar functions, i.e., mapping a single eigenvalue to a single filtered value, thus ignoring the global pattern of the spectrum. Furthermore, these filters are often constructed based on some fixed-order polynomials, which have limited expressiveness and flexibility. To tackle these issues, we introduce Specformer, which effectively encodes the set of all eigenvalues and performs self-attention in the spectral domain, leading to a learnable set-to-set spectral filter. We also design a decoder with learnable bases to enable non-local graph convolution. Importantly, Specformer is equivariant to permutation. By stacking multiple Specformer layers, one can build a powerful spectral GNN. On synthetic datasets, we show that our Specformer can better recover ground-truth spectral filters than other spectral GNNs. Extensive experiments of both node-level and graph-level tasks on real-world graph datasets show that our Specformer outperforms state-of-the-art GNNs and learns meaningful spectrum patterns. Code and data are available at https://github.com/bdy9527/Specformer.

****

MetaP: How to Transfer Your Knowledge on Learning Hidden Physics

Lu Zhang,Huaiqian You,Tian Gao,Mo Yu,Chung-Hao Lee,Yue Yu

Gradient-based meta-learning methods have primarily focused on classical machine learning tasks such as image classification and function regression, where they were found to perform well by recovering the underlying common representation among a set of given tasks. Recently, PDE-solving deep learning methods, such as neural operators, are starting to make an important impact on learning and predicting the response of a complex physical system directly from observational data. Since the data acquisition in this context is commonly challenging and costly, the call of utilization and transfer of existing knowledge to new and unseen physical systems is even more acute.

Herein, we propose a novel meta-learnt approach for transfer-learning knowledge between neural operators, which can be seen as transferring the knowledge of solution operators between governing (unknown) PDEs with varying parameter fields. With the key theoretical observation that the underlying parameter field can be captured in the first layer of the neural operator model, in contrast to typical final-layer transfer in existing meta-learning methods, our approach is a provably universal solution operator for multiple PDE solving tasks. As applications, we demonstrate the efficacy of our proposed approach on heterogeneous material modeling tasks, which shows that our method can handle complex and nonlinear physical response learning tasks while greatly improving the sampling efficiency in new and unseen materials.

****

CommsVAE: Learning the brain's macroscale communication dynamics using coupled sequential VAEs

Eloy Geenjaar,Noah Lewis,Amrit Kashyap,Robyn Miller,Vince Calhoun

Communication within or between complex systems is commonplace in the natural sciences and fields such as graph neural networks. The brain is a perfect example of such a complex system, where communication between brain regions is constantly being orchestrated. To analyze communication, the brain is often split up into anatomical regions that each perform certain computations. These regions must interact and communicate with each other to perform tasks and support higher-level cognition. On a macroscale, these regions communicate through signal propagation along the cortex and along white matter tracts over longer distances. When and what types of signals are communicated over time is an unsolved problem and is often studied using either functional or structural data. In this paper, we propose a non-linear generative approach to communication from functional data. We address three issues with common connectivity approaches by explicitly modeling the directionality of communication, finding communication at each timestep, and encouraging sparsity. To evaluate our model, we simulate temporal data that has sparse communication between nodes embedded in it and show that our model can uncover the expected communication dynamics. Subsequently, we apply our model to temporal neural data from multiple tasks and show that our approach models communication that is more specific to each task. The specificity of our method means it can have an impact on the understanding of psychiatric disorders, which are believed to be related to highly specific communication between brain regions compared to controls. In sum, we propose a general model for dynamic communication learning on graphs, and show its applicability to a subfield of the natural sciences, with potential widespread scientific impact.

****

Answer Me if You Can: Debiasing Video Question Answering via Answering Unanswerable Questions

Dohwan Ko,Heeji Won,Jongha Kim,Miso CHOI,Byungseok Roh,Hyunwoo J. Kim

Video Question Answering (VideoQA) is a task to predict a correct answer given a question-video pair. Recent studies have shown that most VideoQA models rely on spurious correlations induced by various biases when predicting an answer. For instance, VideoQA models tend to predict `two’ as an answer without considering the video if a question starts with ``How many’' since the majority of answers to such type of questions are `two’. In causal inference, such bias ($\textit{question type}$), which simultaneously affects the input X ($\textit{How many...}$) and the answer Y ($\textit{two}$), is referred to as a confounder Z that hinders a model from learning the true relationship between the input and the answer. The effect of the confounders Z can be removed with a causal intervention $P(Y|do(X))$ when Z is observed. However, there exist many unobserved confounders affecting questions and videos, $\textit{e.g.}$, dataset bias induced by annotators who mainly focus on human activities and salient objects resulting in a spurious correlation between videos and questions. To address this problem, we propose a novel framework that learns unobserved confounders by capturing the bias using $\textit{unanswerable}$ questions, which refers to an artificially constructed VQA sample with a video and a question from two different samples, and leverages the confounders for debiasing a VQA model through causal intervention. We demonstrate that our confounders successfully capture the dataset bias by investigating which part in a video or question that confounders pay attention to. Our experiments on multiple VideoQA benchmark datasets show the effectiveness of the proposed debiasing framework, resulting in an even larger performance gap compared to biased models under the distribution shift.

****

Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought

Abulhair Saparov,He He

Large language models (LLMs) have shown remarkable reasoning capabilities given chain-of-thought prompts (examples with intermediate reasoning steps). Existing benchmarks measure reasoning ability indirectly, by evaluating accuracy on downstream tasks such as mathematical reasoning. However, it is unclear how these models obtain the answers and whether they rely on simple heuristics rather than the generated chain-of-thought. To enable systematic exploration of the reasoning ability of LLMs, we present a new synthetic question-answering dataset called PrOntoQA, where each example is generated from a synthetic world model represented in first-order logic. This allows us to parse the generated chain-of-thought into symbolic proofs for formal analysis. Our analysis on InstructGPT and GPT-3 shows that LLMs are quite capable of making correct individual deduction steps, and so are generally capable of reasoning, even in fictional contexts. However, they have difficulty with proof planning: When multiple valid deduction steps are available, they are not able to systematically explore the different options.

****

Approximation ability of Transformer networks for functions with various smoothness of Besov spaces: error analysis and token extraction

Yuki Wada,Shokichi Takakura,Taiji Suzuki

Although Transformer networks outperform various natural language processing tasks, many aspects of their theoretical nature are still unclear. On the other hand, fully connected neural networks have been extensively studied in terms of their approximation and estimation capability where the target function is included in such function classes as H\"older class and Besov class. In this paper, we study the approximation and estimation error of Transformer networks in a setting where the target function takes a fixed-length sentence as an input and belongs to two variants of Besov spaces known as anisotropic Besov and mixed smooth Besov spaces, in which it is shown that Transformer networks can avoid curse of dimensionality. By overcoming the difficulties in limited interactions among tokens, we prove that Transformer networks can accomplish minimax optimal rate. Our result also shows that token-wise parameter sharing in Transformer networks decreases dependence of the network width on the input length. Moreover, we prove that, under suitable situations, Transformer networks dynamically select tokens to pay careful attention to. This phenomenon matches attention mechanism, on which Transformer networks are based. Our analyses strongly support the reason why Transformer networks have outperformed various natural language processing tasks from a theoretical perspective.

****

Clustering Embedding Tables, Without First Learning Them

Henry Ling-Hei Tsang,Thomas Dybdahl Ahle

Machine learning systems use embedding tables to work with categorical features. These tables may get extremely large in modern recommendation systems, and various methods have been suggested to fit them in memory.

Product- and Residual Vector Quantization are some of the most successful methods for table compression. They function by substituting table rows with references to ``codewords'' picked by k-means clustering. Unfortunately, this means that they must first know the table before compressing it, thus they can only save memory at inference time, not training time. Recent work has employed hashing-based approaches to minimize memory usage during training, however the compression obtained is poorer than that achieved by ``post-training'' quantization.

We demonstrate that combining hashing and clustering based algorithms provides the best of both worlds. By first training a hashing-based ``sketch'', then clustering it, and then training the clustered quantization, our method may achieve compression ratios close to those of post-training quantization with the training time memory reductions of hashing-based methods. We prove that this technique works rigorously in the least-square setting.

****

Architecture Matters in Continual Learning

Seyed Iman Mirzadeh,Arslan Chaudhry,Dong Yin,Timothy Nguyen,Razvan Pascanu,Dilan Gorur,Mehrdad Farajtabar

A large body of research in continual learning is devoted to overcoming the catastrophic forgetting of neural networks by designing new algorithms that are robust to the distribution shifts. However, the majority of these works are strictly focused on the "algorithmic" part of continual learning for a "fixed neural network architecture", and the implications of using different architectures are not clearly understood. The few existing continual learning methods that expand the model also assume a fixed architecture and develop algorithms that can efficiently use the model throughout the learning experience. In contrast, in this work, we build on existing works that study continual learning from a neural network's architecture perspective and provide new insights into how the architecture choice, for the same learning algorithm, can impact stability-plasticity trade-off resulting in markedly different continual learning performance. We empirically analyze the impact of various architectural components providing best practices and recommendations that can improve the continual learning performance irrespective of the learning algorithm.

****

Machine Learning Force Fields with Data Cost Aware Training

Alexander Bukharin,Tianyi Liu,Shengjie Wang,Simiao Zuo,Weihao Gao,Wen Yan,Tuo Zhao

Machine learning force fields (MLFF) have been proposed to accelerate molecular dynamics (MD) simulation, which finds widespread applications in chemistry and biomedical research. Even for the most data-efficient MLFF models, reaching chemical accuracy can require hundreds of frames of force and energy labels generated by expensive quantum mechanical algorithms, which may scale as $O(n^3)$ to $O(n^7)$, with n being the number of basis functions used and typically proportional to the number of atoms.

To address this issue, we propose a multi-stage computational framework -- ASTEROID, which enjoys low training data generation cost without significantly sacrificing MLFFs' accuracy. Specifically, ASTEROID leverages a combination of both large cheap inaccurate data and small expensive accurate data. The motivation behind ASTEROID is that inaccurate data, though incurring large bias, can help capture the sophisticated structures of the underlying force field. Therefore, we first train a MLFF model on a large amount of inaccurate training data, employing a bias-aware loss function to prevent the model from overfitting the potential bias of the inaccurate training data. We then fine-tune the obtained model using a small amount of accurate training data, which preserves the knowledge learned from the inaccurate training data while significantly improving the model's accuracy. Moreover, we propose a variant of ASTEROID based on score matching for the setting where the inaccurate training data are unlabelled. Extensive experiments on MD simulation datasets show that ASTEROID can significantly reduce data generation costs while improving the accuracy of MLFFs.

****

Covariance Matrix Adaptation MAP-Annealing

Matthew Christopher Fontaine,Stefanos Nikolaidis

Single-objective optimization algorithms search for the single highest-quality solution with respect to an objective. Quality diversity (QD) algorithms, such as Covariance Matrix Adaptation MAP-Elites (CMA-ME), search for a collection of solutions that are both high-quality with respect to an objective and diverse with respect to specified measure functions. However, CMA-ME suffers from three major limitations highlighted by the QD community: prematurely abandoning the objective in favor of exploration, struggling to explore flat objectives, and having poor performance for low-resolution archives. We propose a new quality diversity algorithm, Covariance Matrix Adaptation MAP-Annealing (CMA-MAE), that addresses all three limitations. We provide theoretical justifications for the new algorit hm with respect to each limitation. Our theory informs our experiments, which support the theory and show that CMA-MAE achieves state-of-the-art performance.

****

Learning Rewards and Skills to Follow Commands with a Data Efficient Visual-Audio Representation

Peixin Chang,Shuijing Liu,Tianchen Ji,Neeloy Chakraborty,D Livingston McPherson,Katherine Rose Driggs-Campbell

Based on the recent advancements in representation learning, we propose a novel framework for command-following robots with raw sensor inputs. Previous RL-based methods are either difficult to continuously improve after the deployment or require a large number of new labels during the fine-tuning. Motivated by (self-)supervised contrastive learning literature, we propose a novel representation, named VAR++, that generates an intrinsic reward function for command-following robot tasks by associating images with sound commands. After the robot is deployed in a new domain, the representation can be updated intuitively and data-efficiently by non-expert, and the robots are able to fulfill sound commands without any hand-crafted reward functions. We demonstrate our approach to various sound types and robotic tasks, including navigation and manipulation with raw sensor inputs. In the simulated experiments, we show that our system can continually self-improve in previously unseen scenarios given fewer new labeled data, yet achieves better performance, compared with previous methods.

****

Reinforcement Learning-Based Estimation for Partial Differential Equations

Saviz Mowlavi,Mouhacine Benosman,Saleh Nabi

In systems governed by nonlinear partial differential equations such as fluid flows, the design of state estimators such as Kalman filters relies on a reduced-order model (ROM) that projects the original high-dimensional dynamics onto a computationally tractable low-dimensional space. However, ROMs are prone to large errors, which negatively affects the performance of the estimator. Here, we introduce the reinforcement learning reduced-order estimator (RL-ROE), a ROM-based estimator in which the correction term that takes in the measurements is given by a nonlinear policy trained through reinforcement learning. The nonlinearity of the policy enables the RL-ROE to compensate efficiently for errors of the ROM, while still taking advantage of the imperfect knowledge of the dynamics. Using examples involving the Burgers and Navier-Stokes equations, we show that in the limit of very few sensors, the trained RL-ROE outperforms a Kalman filter designed using the same ROM. Moreover, it yields accurate high-dimensional state estimates for reference trajectories corresponding to various physical parameter values, without direct knowledge of the latter.

****

Heterogeneous-Agent Mirror Learning

Jakub Grudzien Kuba,Xidong Feng,Shiyao Ding,Hao Dong,Yaodong Yang

The necessity for cooperation among intelligent machines has popularised cooperative multi-agent reinforcement learning (MARL) in the artificial intelligence (AI) research community. However, many research endeavours have been focused on developing practical MARL algorithms whose effectiveness has been studied only empirically, thereby lacking theoretical guarantees. As recent studies have revealed, MARL methods often achieve performance that is unstable in terms of reward monotonicity or suboptimal at convergence. To resolve these issues, in this paper, we introduce a novel framework named Heterogeneous-Agent Mirror Learning (HAML) that provides a general template for MARL algorithmic designs. We prove that algorithms derived from the HAML template satisfy the desired properties of the monotonic improvement of the joint reward and the convergence to Nash equilibrium. We verify the practicality of HAML by proving that the current state-of-the-art cooperative MARL algorithms, HATRPO and HAPPO, are in fact HAML instances. Next, as a natural outcome of our theory, we propose HAML extensions of two well-known RL algorithms, HAA2C (for A2C) and HADDPG (for DDPG), and demonstrate their effectiveness against strong baselines on StarCraftII and Multi-Agent MuJoCo task

s.
****

Recursive Time Series Data Augmentation

Amine Mohamed Aboussalah,Minjae Kwon,Raj G Patel,Cheng Chi,Chi-Guhn Lee

Time series observations can be seen as realizations of an underlying dynamical system governed by rules that we typically do not know. For time series learning tasks we create our model using available data. Training on available realizations, where data is limited, often induces severe over-fitting thereby preventing generalization. To address this issue, we introduce a general recursive framework for time series augmentation, which we call the Recursive Interpolation Method (RIM). New augmented time series are generated using a recursive interpolation function from the original time series for use in training. We perform theoretical analysis to characterize the proposed RIM and to guarantee its performance under certain conditions. We apply RIM to diverse synthetic and real-world time series cases to achieve strong performance over non-augmented data on a variety of learning tasks. Our method is also computationally more efficient and leads to better performance when compared to state of the art time series data augmentation.

****

Auto-Encoding Goodness of Fit

Aaron Palmer,Zhiyi Chi,Derek Aguiar,Jinbo Bi

For generative autoencoders to learn a meaningful latent representation for data generation, a careful balance must be achieved between reconstruction error and how close the distribution in the latent space is to the prior. However, this balance is challenging to achieve due to a lack of criteria that work both at the mini-batch (local) and aggregated posterior (global) level. In this work, we develop the Goodness of Fit Autoencoder (GoFAE), which incorporates hypothesis tests at two levels. At the mini-batch level, it uses GoF test statistics as regularization objectives. At a more global level, it selects a regularization coefficient based on higher criticism, i.e., a test on the uniformity of the local GoF p-values. We justify the use of GoF tests by providing a relaxed L_2-Wasserstein bound on the distance between the latent distribution and target prior. We propose to use GoF tests and prove that optimization based on these tests can be done with stochastic gradient (SGD) descent on a compact Riemannian manifold. Empirically, we show that our higher criticism parameter selection procedure balances reconstruction and generation using mutual information and uniformity of p-values respectively. Finally, we show that GoFAE achieves comparable FID scores and mean squared errors with competing deep generative models while retaining statistical indistinguishability from Gaussian in the latent space based on a variety of hypothesis tests.

****

VER: Learning Natural Language Representations for Verbalizing Entities and Relations

Jie Huang,Kevin Chang

Entities and relationships between entities are vital in the real world. Essentially, we understand the world by understanding entities and relations. For instance, to understand a field, e.g., computer science, we need to understand the relevant concepts, e.g., machine learning, and the relationships between concepts, e.g., machine learning and artificial intelligence. To understand a person, we should first know who he/she is and how he/she is related to others. To understand entities and relations, humans may refer to natural language descriptions. For instance, when learning a new scientific term, people usually start by reading its definition in dictionaries or encyclopedias. To know the relationship between two entities, humans tend to create a sentence to connect them. In this paper, we propose VER: A Unified Model for Verbalizing Entities and Relations. Specifically, we attempt to build a system that takes any entity or entity set as input and generates a sentence to represent entities and relations, named ``natural language representation''. Extensive experiments demonstrate that our model can generate high-quality sentences describing entities and entity relationships and facilitate various tasks on entities and relations, including definition modeling, relation modeling, and generative commonsense reasoning.

****

Adaptive IMLE for Few-shot Image Synthesis

Mehran Aghabozorgi,Shichong Peng,Ke Li

Despite their success on large datasets, GANs have been difficult to apply in the few-shot setting, where only a limited number of training examples are provided. Due to mode collapse, GANs tend to ignore some training examples, causing overfitting to a subset of the training dataset, which is small to begin with. A recent method called Implicit Maximum Likelihood Estimation (IMLE) is an alternative to GAN that tries to address this issue. It uses the same kind of generators as GANs but trains it with a different objective that encourages mode coverage. However, the theoretical guarantees of IMLE hold under restrictive conditions, such as the requirement for the optimal likelihood at all data points to be the same. In this paper, we present a more generalized formulation of IMLE which includes the original formulation as a special case, and we prove that the theoretical guarantees hold under weaker conditions. Using this generalized formulation, we further derive a new algorithm, which we dub Adaptive IMLE, which can adapt to the varying difficulty of different training examples. We demonstrate on multiple few-shot image synthesis datasets that our method significantly outperforms existing methods.

****

Understanding the Covariance Structure of Convolutional Filters

Asher Trockman,Devin Willmott,J Zico Kolter

Neural network weights are typically initialized at random from univariate distributions, controlling just the variance of individual weights even in highly-structured operations like convolutions. Recent ViT-inspired convolutional networks such as ConvMixer and ConvNeXt use large-kernel depthwise convolutions whose learned filters have notable structure; this presents an opportunity to study their empirical covariances. In this work, we first observe that such learned filters have highly-structured covariance matrices, and moreover, we find that covariances calculated from small networks may be used to effectively initialize a variety of larger networks of different depths, widths, patch sizes, and kernel sizes, indicating a degree of model-independence to the covariance structure. Motivated by these findings, we then propose a learning-free multivariate initialization scheme for convolutional filters using a simple, closed-form construction of their covariance. Models using our initialization outperform those using traditional univariate initializations, and typically meet or exceed the performance of those initialized from the covariances of learned filters; in some cases, this improvement can be achieved without training the depthwise convolutional filters at all. Our code is available at https://github.com/locuslab/convcov.

****

Reinforcement Logic Rule Learning for Temporal Point Processes

Chao Yang,Lu Wang,Kun Gao,Shuang Li

We aim to learn a set of temporal logic rules to explain the occurrence of temporal events. Leveraging the temporal point process modeling and learning framework, the rule content and rule weights are jointly learned by maximizing the likelihood of the observed noisy event sequences. The proposed algorithm alternates between a master problem, where the rule weights are updated, and a subproblem, where a new rule is searched and included. The formulated master problem is convex and relatively easy to solve, whereas the subproblem requires searching the huge combinatorial rule predicate and relationship space. To tackle this challenge, we propose a neural search policy to learn to generate the new rule content as a sequence of actions. The policy parameters will be trained end-to-end using the reinforcement learning framework, where the reward signals can be efficiently queried by evaluating the subproblem objective. The trained policy can be used to generate new rules, and moreover, the well-trained policies can be directly transferred to other tasks to speed up the rule searching procedure in the new task. We evaluate our methods on both synthetic and real-world datasets, obtaining promising results.

****

Masked Distillation with Receptive Tokens

Tao Huang,Yuan Zhang,Shan You,Fei Wang,Chen Qian,Jian Cao,Chang Xu

Distilling from the feature maps can be fairly effective for dense prediction tasks since both the feature discriminability and localization information can be well transferred. However, not every pixel contributes equally to the performance, and a good student should learn from what really matters to the teacher. In this paper, we introduce a learnable embedding dubbed receptive token to locate the pixels of interests (PoIs) in the feature map, with a distillation mask generated via pixel-wise attention. Then the masked distillation will be performed via the pixel-wise reconstruction. In this way, a distillation mask refers to a pattern of pixel dependencies. We thus adopt multiple receptive tokens to investigate more sophisticated and informative pixel dependencies within feature maps to enhance the distillation. To obtain a group of masks, the receptive tokens are learned via the regular task loss but with teacher fixed, and we also leverage a Dice loss to enrich the diversity of obtained masks. Our method dubbed MasKD is simple and practical, and needs no priors of ground-truth labels, which can apply to various dense prediction tasks. Experiments show that our MasKD can achieve state-of-the-art performance consistently on object detection and semantic segmentation benchmarks.

****

Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms

Linbo Liu,Youngsuk Park,Trong Nghia Hoang,Hilaf Hasson,Luke Huan

This work studies the threats of adversarial attack on multivariate probabilistic forecasting models and viable defense mechanisms. Our studies discover a new attack pattern that negatively impact the forecasting of a target time series via making strategic, sparse (imperceptible) modifications to the past observations of a small number of other time series. To mitigate the impact of such attack, we have developed two defense strategies. First, we extend a previously developed randomized smoothing technique in classification to multivariate forecasting scenarios. Second, we develop an adversarial training algorithm that learns to create adversarial examples and at the same time optimizes the forecasting model to improve its robustness against such adversarial simulation. Extensive experiments on real-world datasets confirm that our attack schemes are powerful and our defense algorithms are more effective compared with baseline defense mechanisms.

****

TextShield: Beyond Successfully Detecting Adversarial Sentences in text classification

Lingfeng Shen,Ze Zhang,Haiyun Jiang,Ying Chen

Adversarial attack serves as a major challenge for neural network models in NLP, which precludes the model's deployment in safety-critical applications. A recent line of work, detection-based defense, aims to distinguish adversarial sentences from benign ones. However, {the core limitation of previous detection methods is being incapable of giving correct predictions on adversarial sentences unlike defense methods from other paradigms.} To solve this issue, this paper proposes TextShield: (1) we discover a link between text attack and saliency information, and then we propose a saliency-based detector, which can effectively detect whether an input sentence is adversarial or not. (2) We design a saliency-based corrector, which converts the detected adversary sentences to benign ones. By combining the saliency-based detector and corrector, TextShield extends the detection-only paradigm to a detection-correction paradigm, thus filling the gap in the existing detection-based defense. Comprehensive experiments show that (a) TextShield consistently achieves higher or comparable performance than state-of-the-art defense methods across various attacks on different benchmarks. (b) our saliency-based detector outperforms existing detectors for detecting adversarial sentences.

****

Efficient Deep Reinforcement Learning Requires Regulating Overfitting

Qiyang Li,Aviral Kumar,Ilya Kostrikov,Sergey Levine

Deep reinforcement learning algorithms that learn policies by trial-and-error must learn from limited amounts of data collected by actively interacting with the environment. While many prior works have shown that proper regularization techniques are crucial for enabling data-efficient RL, a general understanding of the bottlenecks in data-efficient RL has remained unclear. Consequently, it has been difficult to devise a universal technique that works well across all domains. In this paper, we attempt to understand the primary bottleneck in sample-efficient deep RL by examining several potential hypotheses such as non-stationarity, excessive action distribution shift, and overfitting. We perform thorough empirical analysis on state-based DeepMind control suite (DMC) tasks in a controlled and systematic way to show that high temporal-difference (TD) error on the validation set of transitions is the main culprit that severely affects the performance of deep RL algorithms, and prior methods that lead to good performance do in fact, control the validation TD error to be low. This observation gives us a robust principle for making deep RL efficient: we can hill-climb on the validation TD error by utilizing any form of regularization techniques from supervised learning. We show that a simple online model selection method that targets the validation TD error is effective across state-based DMC and Gym tasks.

****

Nuisances via Negativa: Adjusting for Spurious Correlations via Data Augmentation

Aahlad Manas Puli,Nitish Joshi,He He,Rajesh Ranganath

There exist features that are related to the label in the same way across different settings for that task; these are semantic features or semantics. Features with varying relationships to the label are nuisances. For example, in detecting cows from natural images, the shape of the head is a semantic and because images of cows often have grass backgrounds but not always, the background is a nuisance. Relationships between a nuisance and the label are unstable across settings and, consequently, models that exploit nuisance-label relationships face performance degradation when these relationships change. Direct knowledge of a nuisance helps build models that are robust to such changes, but knowledge of a nuisance requires extra annotations beyond the label and the covariates. In this paper, we develop an alternative way to produces robust models by data augmentation. These data augmentations corrupt semantic information to produce models that identify and adjust for where nuisances drive predictions. We study semantic corruptions in powering different robust-modeling methods for multiple out-of distribution (OOD) tasks like classifying waterbirds, natural language inference, and detecting Cardiomegaly in chest X-rays.

****

GNN Domain Adaptation using Optimal Transport

Qi Zhu,Yizhu Jiao,Haonan Wang,Natalia Ponomareva,Bryan Perozzi

While Graph Convolutional Networks (GCNs) have recently grown in popularity due to their excellent performance on graph data, their performance under domain shift has not been studied extensively. In this work, we first explore the ability of GCNs to generalize to out-of-distribution data using contextual stochastic block models (CSBMs) on the node classification task. Our results in this area provide the first generalization criteria for GCNs on feature distribution and structure changes. Next we examine a popular Unsupervised Domain Adaptation (UDA) covariate shift assumption and demonstrate that it rarely holds for graph data. Motivated by these results, we propose addressing bias in graph models using domain adaptation with optimal transport - GDOT which features a transportation plan that minimizes the cost of the joint feature and estimated label distribution $P(X,\hat{Y})$ between source and target domains. Additionally, we demonstrate that such transportation cost metric serves as a good proxy for estimating transferability between source and target graphs, and is better as a transferability metric than other common metrics like maximum mean discrepancy (MMD). In our controlled CSBM experiments, GDOT demonstrates robustness towards distributional shift, resulting in 90\% ROC AUC (vs.\ the second-best algorithm achieving <80\% on feature shift). Comprehensive experiments on both semi-supervised and supervis ed real-world node classification problems show that our method is the only one that performs consistently better than baseline GNNs in the cross-domain adaptation setting.

****

Ask Me Anything: A simple strategy for prompting language models

Simran Arora,Avanika Narayan,Mayee F Chen,Laurel Orr,Neel Guha,Kush Bhatia,Ines Chami,Christopher Re

Large language models (LLMs) transfer well to new tasks out-of-the-box simply given a natural language prompt that demonstrates how to perform the task and no additional training. Prompting is a brittle process wherein small modifications to the prompt can cause large variations in the model predictions, and therefore significant effort is dedicated towards designing a painstakingly crafted "perfect prompt" for a task. To mitigate the high degree of effort, we instead ask whether collecting multiple decent, yet imperfect, prompts and aggregating them can lead to a high quality prompting strategy. Our observations motivate our proposed method, Ask Me Anything (AMA). We first develop an understanding of the effective prompt formats, finding question-answering (QA) prompts, which encourage open-ended generation ("Who went to the park?") tend to outperform those that restrict the model outputs ("John went to the park. True or False?"). AMA recursively uses the LLM to transform task inputs to the effective QA format. AM generates multiple questions per input and applies these prompts to collect several noisy "votes" for the input's true label. We find the prompts have varying accuracies and dependencies and thus propose to use weak supervision, a procedure for combining the noisy predictions, to produce the final predictions. We evaluate AMA across open-source model families (EleutherAI, BLOOM, OPT, and T0) and sizes (125M-175B parameters), demonstrating an average performance lift of 10.2\% over the few-shot baseline. This simple strategy enables the open-source GPT-J-6B model to match and exceed the performance of few-shot GPT3-175B on 15 of 20 popular benchmarks. Averaged across these tasks, the GPT-J-6B model outperforms few-shot GPT3-175B. We release our code here: https://github.com/HazyResearch/ama_prompting.

****

Limits of Algorithmic Stability for Distributional Generalization

Neha Hulkund,Vinith Menon Suriyakumar,Taylor W. Killian,Marzyeh Ghassemi

As machine learning models become widely considered in safety critical settings, it is important to understand when models may fail after deployment. One cause of model failure is distribution shift, where the training and test data distributions differ. In this paper we investigate the benefits of training models using methods which are algorithmically stable towards improving model robustness, motivated by recent theoretical developments which show a connection between the two. We use techniques from differentially private stochastic gradient descent (DP-SGD) to control the level of algorithmic stability during training. We compare the performance of algorithmically stable training procedures to stochastic gradient descent (SGD) across a variety of possible distribution shifts - specifically covariate, label, and subpopulation shifts. We find that models trained with algorithmically stable procedures result in models with consistently lower generalization gap across various types of shifts and shift severities. as well as a higher absolute test performance in label shift. Finally, we demonstrate that there is there is a tradeoff between distributional robustness, stability, and performance.

****

WikiWhy: Answering and Explaining Cause-and-Effect Questions

Matthew Ho,Aditya Sharma,Justin Chang,Michael Saxon,Sharon Levy,Yujie Lu,William Yang Wang

As large language models (LLMs) grow larger and more sophisticated, assessing their "reasoning" capabilities in natural language grows more challenging. Recent question answering (QA) benchmarks that attempt to assess reasoning are often limited by a narrow scope of covered situations and subject matters. We introduce WikiWhy, a QA dataset built around a novel auxiliary task: explaining why an answer is true in natural language. WikiWhy contains over 9,000 "why" question-answ er-rationale triples, grounded on Wikipedia facts across a diverse set of topics. Each rationale is a set of supporting statements connecting the question to the answer. WikiWhy serves as a benchmark for the reasoning capabilities of LLMs because it demands rigorous explicit rationales for each answer to demonstrate the acquisition of implicit commonsense knowledge, which is unlikely to be easily memorized. GPT-3 baselines achieve only 38.7% human-evaluated correctness in the end-to-end answer & explain condition, leaving significant room for future improvements.

****

Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient

Ming Yin,Mengdi Wang,Yu-Xiang Wang

Offline reinforcement learning, which aims at optimizing sequential decision-making strategies with historical data, has been extensively applied in real-life applications. State-Of-The-Art algorithms usually leverage powerful function approximators (e.g. neural networks) to alleviate the sample complexity hurdle for better empirical performances. Despite the successes, a more systematic under- standing of the statistical complexity for function approximation remains lacking. Towards bridging the gap, we take a step by considering offline reinforcement learning with differentiable function class approximation (DFA). This function class naturally incorporates a wide range of models with nonlinear/nonconvex structures. We show offline RL with differentiable function approximation is provably efficient by analyzing the pessimistic fitted Q-learning (PFQL) algorithm, and our results provide the theoretical basis for understanding a variety of practical heuristics that rely on Fitted Q-Iteration style design. In addition, we further im- prove our guarantee with a tighter instance-dependent characterization. We hope our work could draw interest in studying reinforcement learning with differentiable function approximation beyond the scope of current research.

****

Proto-Value Networks: Scaling Representation Learning with Auxiliary Tasks

Jesse Farebrother,Joshua Greaves,Rishabh Agarwal,Charline Le Lan,Ross Goroshin,Pablo Samuel Castro,Marc G Bellemare

Auxiliary tasks improve the representations learned by deep reinforcement learning agents. Analytically, their effect is reasonably well-understood; in practice, how-ever, their primary use remains in support of a main learning objective, rather than as a method for learning representations. This is perhaps surprising given that many auxiliary tasks are defined procedurally, and hence can be treated as an essentially infinite source of information about the environment. Based on this observation, we study the effectiveness of auxiliary tasks for learning rich representations, focusing on the setting where the number of tasks and the size of the agent’s network are simultaneously increased. For this purpose, we derive a new family of auxiliary tasks based on the successor measure. These tasks are easy to implement and have appealing theoretical properties. Combined with a suitable off-policy learning rule, the result is a representation learning algorithm that can be understood as extending Mahadevan & Maggioni (2007)’s proto-value functions to deep reinforcement learning – accordingly, we call the resulting object proto-value networks. Through a series of experiments on the Arcade Learning Environment, we demonstrate that proto-value networks produce rich features that may be used to obtain performance comparable to established algorithms, using only linear approximation and a small number (~4M) of interactions with the environment’s reward function.

****

Pseudometric guided online query and update for offline reinforcement learning

Haoran Li,Yang Weng

Offline Reinforcement Learning (RL) extracts effective policies from historical data without the need to interact with the environment. However, the learned policy often suffers large generalization errors in the online environment due to the distributional shift. While existing work mostly focuses on learning a generalizable policy, we propose to adapt the learned policy to fit the online environ ment with limited queries. The goals include querying reasonable actions with limited chances and efficiently modifying the policy. Our insight is to unify these two goals via a proper pseudometric. Intuitively, the metric can compare online and offline states to infer optimal query actions. Additionally, efficient policy updates require good knowledge of the similarity between query results and historical data. Therefore, we propose a unified framework, denoted Pseudometric Guided Offline-to-Online RL (PGO2). Specifically, in deep Q learning, PGO2 has a structural design between the Q-neural network and the Siamese network, which guarantees simultaneous Q-network updating and pseudometric learning, promoting Q-network fine-tuning. In the inference phase, PGO2 solves convex optimizations to identify optimal query actions. We also show that PGO2 training converges to the so-called bisimulation metric with strong theoretical guarantees. Finally, we demonstrate the superiority of PGO2 on diversified datasets.

****

Efficient Data Subset Selection to Generalize Training Across Models: Transductive and Inductive Networks

Eeshaan Jain,Tushar Nandy,Gaurav Aggarwal,Ashish V. Tendulkar,Rishabh K Iyer,Abir De

Subset selection, in recent times, has emerged as a successful approach toward efficient training of models by significantly reducing the amount of data and computational resources required. However, existing methods employ discrete combinatorial and model-specific approaches which lack generalizability--- for each new model, the algorithm has to be executed from the beginning. Therefore, for data subset selection for an unseen architecture, one cannot use the subset chosen for a different model. In this work, we propose SubSelNet, a non-adaptive subset selection framework, which tackles these problems with two main components. First, we introduce an attention-based neural gadget that leverages the graph structure of architectures and acts as a surrogate to trained deep neural networks for quick model prediction. Then, we use these predictions to build subset samplers. This leads us to develop two variants of SubSelNet. The first variant is transductive (called as Transductive-SubSelNet) which computes the subset separately for each model by solving a small optimization problem. Such an optimization is still super fast, thanks to the replacement of explicit model training by the model approximator. The second variant is inductive (called as Inductive-SubSelNet) which computes the subset using a trained subset selector, without any optimization. Most state-of-the-art data subset selection approaches are adaptive, in that the subset selection adapts as the training progresses, and as a result, they require access to the entire data at training time. Our approach, in contrast, is non-adaptive and does the subset selection only once in the beginning, thereby achieving resource and memory efficiency along with compute-efficiency at training time. Our experiments show that both transductive and inductive variants of our models outperform several methods on the quality of the subset chosen and further demonstrate that our method can be used for choosing the best architecture from a set of architectures.

****

Probe Into Multi-agent Adversarial Reinforcement Learning through Mean-Field Optimal Control

Ziming Wang,Fengxiang He,Bohan Wang,Die Gan,Dacheng Tao

Multi-agent adversarial reinforcement learning (MaARL) has shown promise in solving adversarial games. However, the theoretical tools for MaARL's analysis is still elusive. In this paper, we take the first step to theoretically understanding MaARL through mean-field optimal control. Specifically, we model MaARL as a mean-field quantitative differential game between two dynamical systems with implicit terminal constraints. Based on the game, we respectively study the optimal solution and the generalization of the fore-mentioned game. First of all, a two-sided extremism principle (TSEP) is then established as a necessary condition for the optimal solution of the game. We further show that TSEP is also sufficient given that the terminal time is sufficiently small. Secondly, based on the TSEP, a generalization bound for MaARL is proposed. The bound does not explicitly rel

y on the dimensions, norms, or other capacity measures of the model, which are usually prohibitively large in deep learning.
****

Robust Algorithms on Adaptive Inputs from Bounded Adversaries

Yeshwanth Cherapanamjeri,Sandeep Silwal,David Woodruff,Fred Zhang,Qiuyi Zhang,Samson Zhou

We study dynamic algorithms robust to adaptive input generated from sources with bounded capabilities, such as sparsity or limited interaction. For example, we consider robust linear algebraic algorithms when the updates to the input are sparse but given by an adversary with access to a query oracle. We also study robust algorithms in the standard centralized setting, where an adversary queries an algorithm in an adaptive manner, but the number of interactions between the adversary and the algorithm is bounded. We first recall a unified framework of (Hassidim et al., 2020; Beimel et al., 2022; Attias et al., 2023) for answering Q adaptive queries that incurs $\widetilde{\mathcal{O}}(\sqrt{Q})$ overhead in space, which is roughly a quadratic improvement over the na\"{i}ve implementation, and only incurs a logarithmic overhead in query time. Although the general framework has diverse applications in machine learning and data science, such as adaptive distance estimation, kernel density estimation, linear regression, range queries, point queries, and serves as a preliminary benchmark, we demonstrate even better algorithmic improvements for (1) reducing the pre-processing time for adaptive distance estimation and (2) permitting an unlimited number of adaptive queries for kernel density estimation. Finally, we complement our theoretical results with additional empirical evaluations.
****

Chasing All-Round Graph Representation Robustness: Model, Training, and Optimization

Chunhui Zhang,Yijun Tian,Mingxuan Ju,Zheyuan Liu,Yanfang Ye,Nitesh Chawla,Chuxu Zhang

Graph Neural Networks (GNNs) have achieved state-of-the-art results on a variety of graph learning tasks, however, it has been demonstrated that they are vulnerable to adversarial attacks, raising serious security concerns. A lot of studies have been developed to train GNNs in a noisy environment and increase their robustness against adversarial attacks. However, existing methods have not uncovered a principled difficulty: the convoluted mixture distribution between clean and attacked data samples, which leads to sub-optimal model design and limits their frameworks’ robustness. In this work, we first begin by identifying the root cause of mixture distribution, then, for tackling it, we propose a novel method GAME - Graph Adversarial Mixture of Experts to enlarge the model capacity and enrich the representation diversity of adversarial samples, from three perspectives of model, training, and optimization. Specifically, we first propose a plug-and-play GAME layer that can be easily incorporated into any GNNs and enhance their adversarial learning capabilities. Second, we design a decoupling-based graph adversarial training in which the component of the model used to generate adversarial graphs is separated from the component used to update weights. Third, we introduce a graph diversity regularization that enables the model to learn diverse representation and further improves model performance. Extensive experiments demonstrate the effectiveness and advantages of GAME over the state-of-the-art adversarial training methods across various datasets given different attacks.
****

Training Neural Networks with Low-Precision Model Memory

Bingrui Li,Ziteng Wang,Jianfei Chen,Jun Zhu

The demand for memory to store model-related statistics ("model memory") is a major bottleneck for training large neural networks. A promising solution is low-precision optimizers, which reduce the numerical precision of the model memory. However, existing work only compresses the momentum, resulting in suboptimal memory efficiency. This paper proposes Low-Precision Model Memory (LPMM), an optimization framework with the entire model memory kept in low precision. LPMM compresses not only the momentum but also model parameters and gradient accumulators. We identify arithmetic underflow as the main problem in building low-precision op timizers and propose a stochastic quantization method and a microbatching technique to overcome this problem. We analyze the convergence behavior of LPMM and theoretically show how the proposed techniques could affect underflowing, which in turn affects the convergence. We apply LPMM to the SGD optimizer with momentum (SGDM). On several realistic benchmarks, LPMM-SGDM can train neural networks with negligible loss of accuracy while reducing over 70% of the model memory compared to the full-precision SGDM.

****

Raisin: Residual Algorithms for Versatile Offline Reinforcement Learning

Braham Snyder,Yuke Zhu

The residual gradient algorithm (RG), gradient descent of the Mean Squared Bellman Error, brings robust convergence guarantees to bootstrapped value estimation. Meanwhile, the far more common semi-gradient algorithm (SG) suffers from well-known instabilities and divergence. Unfortunately, RG often converges slowly in practice. Baird (1995) proposed residual algorithms (RA), weighted averaging of RG and SG, to combine RG's robust convergence and SG's speed. RA works moderately well in the online setting. We find, however, that RA works disproportionately well in the offline setting. Concretely, we find that merely adding a variable residual component to SAC increases its score on D4RL gym tasks by a median factor of 54. We further show that using the minimum of ten critics lets our algorithm match SAC-N's state-of-the-art returns using 50$\times$ less compute and no additional hyperparameters. In contrast, TD3+BC with the same minimum-of-ten-critics trick does not match SAC-N's returns on a handful of environments.

****

VQR: Automated Software Vulnerability Repair Through Vulnerability Queries

Michael Fu,Van Nguyen,Chakkrit Tantithamthavorn,Trung Le,Dinh Phung

Recently, automated vulnerability repair (AVR) approaches have been widely adopted to combat the increasing number of software security issues. In particular, transformer-based models achieve competitive results. While existing models are learned to generate vulnerability repairs, existing AVR models lack a mechanism to provide their models with the precise location of vulnerable code (i.e., models may generate repairs for the non-vulnerable areas). To address this problem, we base our framework on the VIT-based approaches for object detection that learn to locate bounding boxes via the cross-matching between object queries and image patches. We cross-match vulnerability queries and their corresponding vulnerable code areas through the cross-attention mechanism to generate more accurate repairs. To strengthen our cross-matching, we propose to learn a novel vulnerability query mask that greatly focuses on vulnerable code areas and integrate it into the cross-attention. Moreover, we also incorporate the vulnerability query mask into the self-attention to learn embeddings that emphasize the vulnerable areas of a program. Through an extensive evaluation using the real-world 5,417 vulnerabilities, our approach outperforms all of the baseline methods by 3.39%-33.21%. The training code and pre-trained models are available at https://github.com/AVR-VQR/VQR.

****

Fully Online Meta Learning

Jathushan Rajasegaran,Chelsea Finn,Sergey Levine

While deep networks can learn complex functions such as classifiers, detectors, and trackers, many applications require models that continually adapt to changing input distributions, changing tasks, and changing environmental conditions. Indeed, this ability to continuously accrue knowledge and use past experience to learn new tasks quickly in continual settings is one of the key properties of an intelligent system. For complex and high-dimensional problems, simply updating the model continually with standard learning algorithms such as gradient descent may result in slow adaptation. Meta-learning can provide a powerful tool to accelerate adaptation yet is conventionally studied in batch settings. In this paper, we study how meta-learning can be applied to tackle online problems of this nature, simultaneously adapting to changing tasks and input distributions and meta-training the model in order to adapt more quickly in the future. Extending meta-learning into the online setting presents its own challenges, and although seve ral prior methods have studied related problems, they generally require a discrete notion of tasks, with known ground-truth task boundaries. Such methods typically adapt to each task in sequence, resetting the model between tasks, rather than adapting continuously across tasks. In many real-world settings, such discrete boundaries are unavailable, and may not even exist. To address these settings, we propose a Fully Online Meta-Learning (FOML) algorithm, which does not require any ground truth knowledge about the task boundaries and stays fully online without resetting back to pre-trained weights. Our experiments show that FOML was able to learn new tasks faster than the state-of-the-art online learning methods on Rainbow-MNIST, CIFAR100 and CELEBA datasets.

****

Learning Globally Smooth Functions on Manifolds

Juan Cervino,Luiz F. O. Chamon,Benjamin David Haeffele,Rene Vidal,Alejandro Ribeiro

Smoothness and low dimensional structures play central roles in improving generalization and stability in learning and statistics. The combination of these properties has led to many advances in semi-supervised learning, generative modeling, and control of dynamical systems. However, learning smooth functions is generally challenging, except in simple cases such as learning linear or kernel models. Typical methods are either too conservative, relying on crude upper bounds such as spectral normalization, too lax, penalizing smoothness on average, or too computationally intensive, requiring the solution of large-scale semi-definite programs. These issues are only exacerbated when trying to simultaneously exploit low dimensionality using, e.g., manifolds. This work proposes to overcome these obstacles by combining techniques from semi-infinite constrained learning and manifold regularization. To do so, it shows that, under typical conditions, the problem of learning a Lipschitz continuous function on a manifold is equivalent to a dynamically weighted manifold regularization problem. This observation leads to a practical algorithm based on a weighted Laplacian penalty whose weights are adapted using stochastic gradient techniques. We prove that, under mild conditions, this method estimates the Lipschitz constant of the solution, learning a globally smooth solution as a byproduct. Numerical examples illustrate the advantages of using this method to impose global smoothness on manifolds as opposed to imposing smoothness on average.

****

On Representing Mixed-Integer Linear Programs by Graph Neural Networks

Ziang Chen,Jialin Liu,Xinshang Wang,Wotao Yin

While Mixed-integer linear programming (MILP) is NP-hard in general, practical MILP has received roughly 100--fold speedup in the past twenty years. Still, many classes of MILPs quickly become unsolvable as their sizes increase, motivating researchers to seek new acceleration techniques for MILPs. With deep learning, they have obtained strong empirical results, and many results were obtained by applying graph neural networks (GNNs) to making decisions in various stages of MILP solution processes. This work discovers a fundamental limitation: there exist feasible and infeasible MILPs that all GNNs will, however, treat equally, indicating GNN's lacking power to express general MILPs. Then, we show that, by restricting the MILPs to unfoldable ones or by adding random features, there exist GNNs that can reliably predict MILP feasibility, optimal objective values, and optimal solutions up to prescribed precision. We conducted small-scale numerical experiments to validate our theoretical findings.

****

LEARNING DYNAMIC ABSTRACT REPRESENTATIONS FOR SAMPLE-EFFICIENT REINFORCEMENT LEARNING

Mehdi Dadvar,Rashmeet Kaur Nayyar,Siddharth Srivastava

In many real-world problems, the learning agent needs to learn a problem’s abstractions and solution simultaneously. However, most such abstractions need to be designed and refined by hand for different problems and domains of application. This paper presents a novel top-down approach for constructing state abstractions while carrying out reinforcement learning. Starting with state variables and a simulator, it presents a novel domain-independent approach for dynamically comp uting an abstraction based on the dispersion of Q-values in abstract states as the agent continues acting and learning. Extensive empirical evaluation on multiple domains and problems shows that this approach automatically learns abstractions that are finely-tuned to the problem, yield powerful sample efficiency, and result in the RL agent significantly outperforming existing approaches.

****

Fighting Fire with Fire: Contrastive Debiasing without Bias-free Data via Generative Bias-transformation

Yeonsung Jung,Hajin Shim,June Yong Yang,Eunho Yang

Despite their remarkable ability to generalize with over-capacity networks, deep neural networks often abuse bias instead of using the actual task-related information for discriminative tasks. Since such shortcuts are only effective within the collected dataset, the resulting biased model underperforms on real-world inputs. To counteract the influence of bias, existing methods either exploit auxiliary information which is rarely obtainable in practice, or sift bias-free samples to exploit them for debiasing. However, such presumptions about the availability of the auxiliary information or bias-free samples are not always guaranteed and the existing methods could break down due to the unmet presumptions. In this paper, we propose Contrastive Debiasing via Generative Bias-transformation (CDvG) which is capable of operating without exploiting bias labels and bias-free samples explicitly. Motivated by our observation that not only discriminative models but also image translation models tend to focus on the easy-to-learn bias, CDvG employs a image translation model to transform the bias to another mode of bias while preserving task-relevant information. Through contrastive learning, we set transformed biased views against another, learning bias-invariant representations. Especially, as the bias has a stronger correlation or is easier to perceive compared to the signal, the translation model is more likely to be a bias translation model, resulting in better debiasing effect. Experimental results demonstrate that CDvG outperforms the state-of-the-arts, especially when bias-free samples are extremely scarce.

****

On Representing Linear Programs by Graph Neural Networks

Ziang Chen,Jialin Liu,Xinshang Wang,Wotao Yin

Learning to optimize is a rapidly growing area that aims to solve optimization problems or improve existing optimization algorithms using machine learning (ML). In particular, the graph neural network (GNN) is considered a suitable ML model for optimization problems whose variables and constraints are permutation--invariant, for example, the linear program (LP). While the literature has reported encouraging numerical results, this paper establishes the theoretical foundation of applying GNNs to solving LPs. Given any size limit of LPs, we construct a GNN that maps different LPs to different outputs. We show that properly built GNNs can reliably predict feasibility, boundedness, and an optimal solution for each LP in a broad class. Our proofs are based upon the recently--discovered connections between the Weisfeiler--Lehman isomorphism test and the GNN. To validate our results, we train a simple GNN and present its accuracy in mapping LPs to their feasibilities and solutions.

****

On the Importance and Applicability of Pre-Training for Federated Learning

Hong-You Chen,Cheng-Hao Tu,Ziwei Li,Han Wei Shen,Wei-Lun Chao

Pre-training is prevalent in nowadays deep learning to improve the learned model's performance. However, in the literature on federated learning (FL), neural networks are mostly initialized with random weights. These attract our interest in conducting a systematic study to explore pre-training for FL. Across multiple visual recognition benchmarks, we found that pre-training can not only improve FL, but also close its accuracy gap to the counterpart centralized learning, especially in the challenging cases of non-IID clients' data. To make our findings applicable to situations where pre-trained models are not directly available, we explore pre-training with synthetic data or even with clients' data in a decentralized manner, and found that they can already improve FL notably. Interestingly, many of the techniques we explore are complementary to each other to further bo ost the performance, and we view this as a critical result toward scaling up deep FL for real-world applications. We conclude our paper with an attempt to understand the effect of pre-training on FL. We found that pre-training enables the learned global models under different clients' data conditions to converge to the same loss basin, and makes global aggregation in FL more stable. Nevertheless, pre-training seems to not alleviate local model drifting, a fundamental problem in FL under non-IID data.

****

Scale-invariant Bayesian Neural Networks with Connectivity Tangent Kernel

SungYub Kim,Sihwan Park,Kyung-Su Kim,Eunho Yang

Studying the loss landscapes of neural networks is critical to identifying generalizations and avoiding overconfident predictions. Flatness, which measures the perturbation resilience of pre-trained parameters for loss values, is widely acknowledged as an essential predictor of generalization. While the concept of flatness has been formalized as a PAC-Bayes bound, it has been observed that the generalization bounds can vary arbitrarily depending on the scale of the model parameters. Despite previous attempts to address this issue, generalization bounds remain vulnerable to function-preserving scaling transformations or are limited to impractical network structures. In this paper, we introduce new PAC-Bayes prior and posterior distributions invariant to scaling transformations, achieved through the \textit{decomposition of perturbations into scale and connectivity components}. In this way, this approach expands the range of networks to which the resulting generalization bound can be applied, including those with practical transformations such as weight decay with batch normalization. Moreover, we demonstrate that scale-dependency issues of flatness can adversely affect the uncertainty calibration of Laplace approximation, and we propose a solution using our invariant posterior. Our proposed invariant posterior allows for effective measurement of flatness and calibration with low complexity while remaining invariant to practical parameter transformations, also applying it as a reliable predictor of neural network generalization.

****

Autoregressive Graph Network for Learning Multi-step Physics

Sakthi Kumar Arul Prakash,Conrad Tucker

In this work, we propose an Autoregressive Graph Network~(AGN) that learns forward physics using a temporal inductive bias. Currently, temporal state space information is provided as additional input to a GN when generating roll-out physics simulations. While this relatively increases the network's predictive performance over multiple time steps, a temporal model enables the network to induce and learn temporal biases. In dynamical systems, the arrow of time simplifies possible interactions in the sense that we can assume current observations to be dependent on preceding states. The autoregressive property naturally induces the arrow of time and can further constrain physics-induced GNs to conserve symmetries over long time-steps. Our proposed GN encodes temporal state information using an autoregressive encoder that can parallelly compute latent temporal embeddings over multiple time steps during a single forward pass. We perform case studies that compare multi-step forward predictions against baseline data-driven GNs across diverse datasets that feature different particle interactions. Our approach outperforms the state of the art GN and physics-induced GNs in 9 out of 10 and in 7 out of 10 particle physics datasets when conditioned on optimal historical states.

****

Simple initialization and parametrization of sinusoidal networks via their kernel bandwidth

Filipe de Avila Belbute-Peres,J Zico Kolter

Neural networks with sinusoidal activations have been proposed as an alternative to networks with traditional activation functions. Despite their promise, particularly for learning implicit models, their training behavior is not yet fully understood, leading to a number of empirical design choices that are not well justified. In this work, we first propose a simplified version of such sinusoidal neural networks, which allows both for easier practical implementation and simple r theoretical analysis. We then analyze the behavior of these networks from the neural tangent kernel perspective and demonstrate that their kernel approximates a low-pass filter with an adjustable bandwidth. Finally, we utilize these insights to inform the sinusoidal network initialization, optimizing their performance for each of a series of tasks, including learning implicit models and solving differential equations.

****

Quasiconvex Shallow Neural Network

CHENHAN XIAO,Yang Weng

Deep neural networks generally have highly non-convex structures, resulting in multiple local optima of network weights. The non-convex network is likely to fail, i.e., being trapped in bad local optima with large errors, especially when the task involves convexity (e.g., linearly separable classification). While convexity is essential in training neural networks, designing a convex network structure without strong assumptions (e.g., linearity) of activation or loss function is challenging. To extract and utilize convexity, this paper presents the QuasiConvex shallow Neural Network (QCNN) architecture with mild assumptions. We first decompose the network into building blocks where quasiconvexity is thoroughly studied. Then, we design additional layers to preserve quasiconvexity where such building blocks are integrated into general networks. The proposed QCNN, interpreted as a quasiconvex optimization problem, allows for efficient training with theoretical guarantees. Specifically, we construct equivalent convex feasibility problems to solve the quasiconvex optimization problem. Our theoretical results are verified via extensive experiments on common machine learning tasks. The quasiconvex structure in QCNN demonstrates even better learning ability than non-convex deep networks in some tasks.

****

The Best of Both Worlds: Accurate Global and Personalized Models through Federated Learning with Data-Free Hyper-Knowledge Distillation

Huancheng Chen,Chaining Wang,Haris Vikalo

Heterogeneity of data distributed across clients limits the performance of global models trained through federated learning, especially in the settings with highly imbalanced class distributions of local datasets. In recent years, personalized federated learning (pFL) has emerged as a potential solution to the challenges presented by heterogeneous data. However, existing pFL methods typically enhance performance of local models at the expense of the global model's accuracy. We propose FedHKD (Federated Hyper-Knowledge Distillation), a novel FL algorithm in which clients rely on knowledge distillation (KD) to train local models. In particular, each client extracts and sends to the server the means of local data representations and the corresponding soft predictions -- information that we refer to as ``hyper-knowledge". The server aggregates this information and broadcasts it to the clients in support of local training. Notably, unlike other KD-based pFL methods, FedHKD does not rely on a public dataset nor it deploys a generative model at the server. We analyze convergence of FedHKD and conduct extensive experiments on visual datasets in a variety of scenarios, demonstrating that FedHKD provides significant improvement in both personalized as well as global model performance compared to state-of-the-art FL methods designed for heterogeneous data settings.

****

Minimalistic Unsupervised Representation Learning with the Sparse Manifold Transform

Yubei Chen,Zeyu Yun,Yi Ma,Bruno Olshausen,Yann LeCun

We describe a minimalistic and interpretable method for unsupervised representation learning that does not require data augmentation, hyperparameter tuning, or other engineering designs, but nonetheless achieves performance close to the state-of-the-art (SOTA) SSL methods. Our approach leverages the sparse manifold transform, which unifies sparse coding, manifold learning, and slow feature analysis. With a one-layer deterministic (one training epoch) sparse manifold transform, it is possible to achieve 99.3% KNN top-1 accuracy on MNIST, 81.1% KNN top-1 accuracy on CIFAR-10, and 53.2% on CIFAR-100. With simple gray-scale aug mentation, the model achieves 83.2% KNN top-1 accuracy on CIFAR-10 and 57% on CIFAR-100. These results significantly close the gap between simplistic ``white-box'' methods and SOTA methods. We also provide visualization to illustrate how an unsupervised representation transform is formed. The proposed method is closely connected to latent-embedding self-supervised methods and can be treated as the simplest form of VICReg. Though a small performance gap remains between our simple constructive model and SOTA methods, the evidence points to this as a promising direction for achieving a principled and white-box approach to unsupervised representation learning, which has potential to significantly improve learning efficiency.

****

Over-Training with Mixup May Hurt Generalization

Zixuan Liu,Ziqiao Wang,Hongyu Guo,Yongyi Mao

Mixup, which creates synthetic training instances by linearly interpolating random sample pairs, is a simple and yet effective regularization technique to boost the performance of deep models trained with SGD. In this work, we report a previously unobserved phenomenon in Mixup raining: on a number of standard datasets, the performance of Mixup-trained models starts to decay after training for a large number of epochs, giving rise to a U-shaped generalization curve. This behavior is further aggravated when the size of original dataset is reduced. To help understand such a behavior of Mixup, we show theoretically that Mixup training may introduce undesired data-dependent label noises to the synthesized data. Via analyzing a least-square regression problem with a random feature model, we explain why noisy labels may cause the U-shaped curve to occur: Mixup improves generalization through fitting the clean patterns at the early training stage, but as training progresses, Mixup becomes over-fitting to the noise in the synthetic data. Extensive experiments are performed on a variety of benchmark datasets, validating this explanation.

****

HiCLIP: Contrastive Language-Image Pretraining with Hierarchy-aware Attention

Shijie Geng,Jianbo Yuan,Yu Tian,Yuxiao Chen,Yongfeng Zhang

The success of large-scale contrastive vision-language pretraining (CLIP) has benefited both visual recognition and multimodal content understanding. The concise design brings CLIP the advantage in inference efficiency against other vision-language models with heavier cross-attention fusion layers, making it a popular choice for a wide spectrum of downstream tasks. However, CLIP does not explicitly capture the hierarchical nature of high-level and fine-grained semantics conveyed in images and texts, which is arguably critical to vision-language understanding and reasoning. To this end, we equip both the visual and language branches in CLIP with hierarchy-aware attentions, namely Hierarchy-aware CLIP (HiCLIP), to progressively discover semantic hierarchies layer-by-layer from both images and texts in an unsupervised manner. As a result, such hierarchical aggregation significantly improves the cross-modal alignment. To demonstrate the advantages of HiCLIP, we conduct qualitative analysis on its unsupervised hierarchy induction during inference, as well as extensive quantitative experiments on both visual recognition and vision-language downstream tasks.

****

Quantile Risk Control: A Flexible Framework for Bounding the Probability of High-Loss Predictions

Jake Snell,Thomas P Zollo,Zhun Deng,Toniann Pitassi,Richard Zemel

Rigorous guarantees about the performance of predictive algorithms are necessary in order to ensure their responsible use. Previous work has largely focused on bounding the expected loss of a predictor, but this is not sufficient in many risk-sensitive applications where the distribution of errors is important. In this work, we propose a flexible framework to produce a family of bounds on quantiles of the loss distribution incurred by a predictor. Our method takes advantage of the order statistics of the observed loss values rather than relying on the sample mean alone. We show that a quantile is an informative way of quantifying predictive performance, and that our framework applies to a variety of quantile-based metrics, each targeting important subsets of the data distribution. We analy ze the theoretical properties of our proposed method and demonstrate its ability to rigorously control loss quantiles on several real-world datasets.

****

Dynamic Neural Network is All You Need: Understanding the Robustness of Dynamic Mechanisms in Neural Networks

Mirazul Haque,Wei Yang

Deep Neural Networks (DNN) based solutions are being used to solve different day-to-day problems. Recently, DNNs are being deployed in real-time systems, and lowering the energy consumption and response time has become the need of the hour. To address this scenario, researchers have proposed early-exit Dynamic Neural Networks (DyNNs), where the computation is dynamic based on the input complexity. DyNNs are generally designed and based on larger static DNNs (SDNN). As the DyNNs decrease the energy consumption, it also becomes important to evaluate the robustness of DyNNs to ensure safety. However, there has not been a significant number of works focusing on the robustness of DyNNs. To address this issue, we propose systematic studies to evaluate the robustness of DyNNs. For that purpose, we propose four research questions. These studies are performed on three models and two datasets. Through the studies, we find that DyNNs are more robust than SDNNs, and DyNNs can be used to generate adversarial samples efficiently. We also provide insight into the design choices through research studies. Finally, we propose a novel attack that can decrease the effectiveness of the DyNNs and can be used to evaluate design choices in DyNN.

****

AutoMoE: Neural Architecture Search for Efficient Sparsely Activated Transformers

Ganesh Jawahar,Subhabrata Mukherjee,Xiaodong Liu,Young Jin Kim,Muhammad Abdul-Mageed,Laks V. S. Lakshmanan,Ahmed Hassan Awadallah,Sebastien Bubeck,Jianfeng Gao

Neural architecture search (NAS) has demonstrated promising results on identifying efficient Transformer architectures which outperform manually designed ones for natural language tasks like neural machine translation (NMT). Existing NAS methods operate on a space of dense architectures, where all of the sub-architecture weights are activated for every input. Motivated by the recent advances in sparsely activated models like the Mixture-of-Experts (MoE) model, we introduce sparse architectures with conditional computation into the NAS search space. Given this expressive search space which subsumes prior densely activated architectures, we develop a new framework AutoMoE to search for efficient sparsely activated sub-Transformers. AutoMoE sparse models obtain (i) 3x FLOPs reduction over manually designed dense Transformers and (ii) 23% FLOPs reduction over state-of-the-art NAS-generated dense sub-Transformers with parity in BLEU score on benchmark datasets for NMT. AutoMoE consists of three training phases: (a) Heterogeneous search space design with dense and sparsely activated Transformer modules (e.g., how many experts? where to place them? what should be their sizes?}; (b) SuperNet training that jointly trains several subnetworks sampled from the large search space by weight-sharing; (c) Evolutionary search for the architecture with the optimal trade-off between task performance and computational constraint like FLOPs and latency.

****

Learning Shareable Bases for Personalized Federated Image Classification

Hong-You Chen,Jike zhong,Mingda Zhang,Xuhui Jia,Hang Qi,Boqing Gong,Wei-Lun Chao,Li Zhang

Personalized federated learning (PFL) aims to leverage the collective wisdom of clients' data while constructing customized models that are tailored to individual client's data distributions. The existing work of PFL mostly aims to personalize for participating clients. In this paper, we focus on a less studied but practically important scenario---generating a personalized model for a new client efficiently. Different from most previous approaches that learn a whole or partial network for each client, we explicitly model the clients' overall meta distribution and embed each client into a low dimension space. We propose FedBasis, a novel PFL algorithm that learns a set of few, shareable basis models, upon which each client only needs to learn the coefficients for combining them into a perso nalized network. FedBasis is parameter-efficient, robust, and more accurate compared to other competitive PFL baselines, especially in a low data regime, without increasing the inference cost. To demonstrate its applicability, we further present a PFL evaluation protocol for image classification, featuring larger data discrepancies across clients in both the image and label spaces as well as more faithful training and test splits.

****

Curriculum-inspired Training for Selective Neural Networks

Rui Liu,Reza Soroushmehr,Barzan Mozafari

We consider the problem of training neural network models for selective classification, where the models have the reject option to abstain from predicting certain examples as needed. Recent advances in curriculum learning have demonstrated the benefit of leveraging the example difficulty scores in training deep neural networks for typical classification settings. Example difficulty scores are even more important in selective classification as a lower prediction error rate can be achieved by rejecting hard examples and accepting easy ones. In this paper, we propose a curriculum-inspired method to train selective neural network models by leveraging example difficulty scores. Our method tailors the curriculum idea to selective neural network training by calibrating the ratio of easy and hard examples in each mini-batch, and exploiting difficulty ordering at the mini-batch level. Our experimental results demonstrate that our method outperforms both the state-of-the-art and alternative methods using vanilla curriculum techniques for training selective neural network models.

****

A Probabilistic Framework For Modular Continual Learning

Lazar Valkov,Akash Srivastava,Dipak Chaudhari,Swarat Chaudhuri,Charles Sutton

Continual learning (CL) algorithms seek to accumulate and transfer knowledge across a sequence of tasks and achieve better performance on each successive task. Modular approaches, which use a different composition of modules for each task and avoid forgetting by design, have been shown to be a promising direction to CL. However, searching through the large space of possible module compositions remains a challenge. In this work, we develop a scalable probabilistic search framework as a solution to this challenge. Our framework has two distinct components. The first is designed to transfer knowledge across similar input domains. To this end, it models each module’s training input distribution and uses a Bayesian model to find the most promising module compositions for a new task. The second component targets transfer across tasks with disparate input distributions or different input spaces and uses Bayesian optimisation to explore the space of module compositions. We show that these two methods can be easily combined and evaluate the resulting approach on two benchmark suites designed to capture different desiderata of CL techniques. The experiments show that our framework offers superior performance compared to state-of-the-art CL baselines.

****

Knowledge-Grounded Reinforcement Learning

Zih-Yun Chiu,Yi-Lin Tuan,William Yang Wang,Michael C. Yip

Receiving knowledge, abiding by laws, and being aware of regulations are common behaviors in human society. Bearing in mind that reinforcement learning (RL) algorithms benefit from mimicking humanity, in this work, we propose that an RL agent can act on external guidance in both its learning process and model deployment, making the agent more socially acceptable. We introduce the concept, Knowledge-Grounded RL (KGRL), with a formal definition that an agent learns to follow external guidelines and develop its own policy. Moving towards the goal of KGRL, we propose a novel actor model with an embedding-based attention mechanism that can attend to either a learnable internal policy or external knowledge. The proposed method is orthogonal to training algorithms, and the external knowledge can be flexibly recomposed, rearranged, and reused in both training and inference stages. Through experiments on tasks with discrete and continuous action space, our KGRL agent is shown to be more sample efficient and generalizable, and it has flexibly rearrangeable knowledge embeddings and interpretable behaviors.

****

Git Re-Basin: Merging Models modulo Permutation Symmetries

Samuel Ainsworth,Jonathan Hayase,Siddhartha Srinivasa

The success of deep learning is due in large part to our ability to solve certain massive non-convex optimization problems with relative ease. Though non-convex optimization is NP-hard, simple algorithms -- often variants of stochastic gradient descent -- exhibit surprising effectiveness in fitting large neural networks in practice. We argue that neural network loss landscapes often contain (nearly) a single basin after accounting for all possible permutation symmetries of hidden units a la Entezari et al. 2021. We introduce three algorithms to permute the units of one model to bring them into alignment with a reference model in order to merge the two models in weight space. This transformation produces a functionally equivalent set of weights that lie in an approximately convex basin near the reference model. Experimentally, we demonstrate the single basin phenomenon across a variety of model architectures and datasets, including the first (to our knowledge) demonstration of zero-barrier linear mode connectivity between independently trained ResNet models on CIFAR-10. Additionally, we identify intriguing phenomena relating model width and training time to mode connectivity. Finally, we discuss shortcomings of the linear mode connectivity hypothesis, including a counterexample to the single basin theory.

****

The Tilted Variational Autoencoder: Improving Out-of-Distribution Detection

Griffin Floto,Stefan Kremer,Mihai Nica

A problem with using the Gaussian distribution as a prior for the variational autoencoder (VAE) is that the set on which Gaussians have high probability density is small as the latent dimension increases. This is an issue because VAEs try to attain both a high likelihood with respect to a prior distribution and at the same time, separation between points for better reconstruction. Therefore, a small volume in the high-density region of the prior is problematic because it restricts the separation of latent points. To ameliorate this, we propose a simple generalization of the Gaussian distribution, called the tilted Gaussian, which has a maximum probability density occurring on a sphere instead of a single point. The tilted Gaussian has exponentially more volume in high-density regions than the standard Gaussian as a function of the distribution dimension. We empirically demonstrate that this simple change in the prior distribution improves VAE performance on the task of detecting unsupervised out-of-distribution (OOD) samples. We also introduce a new OOD testing procedure, called the Will-It-Move test, where the tilted Gaussian achieves remarkable OOD performance.

****

The Role of Coverage in Online Reinforcement Learning

Tengyang Xie,Dylan J Foster,Yu Bai,Nan Jiang,Sham M. Kakade

Coverage conditions---which assert that the data logging distribution adequately covers the state space---play a fundamental role in determining the sample complexity of offline reinforcement learning. While such conditions might seem irrelevant to online reinforcement learning at first glance, we establish a new connection by showing---somewhat surprisingly---that the mere existence of a data distribution with good coverage can enable sample-efficient online RL. Concretely, we show that coverability---that is, existence of a data distribution that satisfies a ubiquitous coverage condition called concentrability---can be viewed as a structural property of the underlying MDP, and can be exploited by standard algorithms for sample-efficient exploration, even when the agent does not know said distribution. We complement this result by proving that several weaker notions of coverage, despite being sufficient for offline RL, are insufficient for online RL. We also show that existing complexity measures for online RL, including Bellman rank and Bellman-Eluder dimension, fail to optimally capture coverability, and propose a new complexity measure, the self-normalized coefficient, to provide a unification.

****

Learning Mixture Models with Simultaneous Data Partitioning and Parameter Estimation

Parth Vipul Sangani,Arjun Shashank Kashettiwar,Durga S,Ganesh Ramakrishnan,Risha bh K Iyer,Abir De

We study a new framework of learning mixture models via data partitioning called PRESTO, wherein we optimize a joint objective function on the model parameters and the partitioning, with each model tailored to perform well on its specific partition. We connect PRESTO to a number of past works in data partitioning, mixture models, and clustering, and show that PRESTO generalizes several loss functions including the k-means and Bregman clustering objective, the Gaussian mixture model objective, mixtures of support vector machines, and mixtures of linear regression. We then propose a new joint discrete-continuous optimization algorithm which achieves a bounded approximation guarantee for any general loss function, thereby achieving guarantees for the afore-mentioned problems as well. We study PRESTO in the context of resource efficient deep learning, where we train smaller resource constrained models on each partition and show that it outperforms existing data partitioning and model pruning/knowledge distillation approaches, which in contrast to PRESTO, require large initial (teacher) models.

****

Estimating Treatment Effects using Neurosymbolic Program Synthesis

Abbavaram Gowtham Reddy,Vineeth N. Balasubramanian

Estimating treatment effects from observational data is a central problem in causal inference. Methods to solve this problem exploit inductive biases and heuristics from causal inference to design multi-head neural network architectures and regularizers. In this work, we propose to use neurosymbolic program synthesis, a data-efficient, and interpretable technique, to solve the treatment effect estimation problem. We theoretically show that neurosymbolic programming can solve the treatment effect estimation problem. By designing a Domain Specific Language (DSL) for treatment effect estimation based on the inductive biases used in literature, we argue that neurosymbolic programming is a better alternative to treatment effect estimation than traditional models. Our empirical study reveals that our model, which implicitly encodes inductive biases in a DSL, achieves better performance on benchmark datasets than the state-of-the-art models.

****

Stateful Active Facilitator: Coordination and Environmental Heterogeneity in Cooperative Multi-Agent Reinforcement Learning

Dianbo Liu,Vedant Shah,Oussama Boussif,Cristian Meo,Anirudh Goyal,Tianmin Shu,Michael Curtis Mozer,Nicolas Heess,Yoshua Bengio

In cooperative multi-agent reinforcement learning, a team of agents works together to achieve a common goal. Different environments or tasks may require varying degrees of coordination among agents in order to achieve the goal in an optimal way. The nature of coordination will depend on properties of the environment—its spatial layout, distribution of obstacles, dynamics, etc. We term this variation of properties within an environment as heterogeneity. Existing literature has not sufficiently addressed the fact that different environments may have different levels of heterogeneity. We formalize the notions of coordination level and heterogeneity level of an environment and present HECOGrid, a suite of multi-agent RL environments that facilitates empirical evaluation of different MARL approaches across different levels of coordination and environmental heterogeneity by providing a quantitative control over coordination and heterogeneity levels of the environment. Further, we propose a Centralized Training Decentralized Execution learning approach called Stateful Active Facilitator (SAF) that enables agents to work efficiently in high-coordination and high-heterogeneity environments through a differentiable and shared knowledge source used during training and dynamic selection from a shared pool of policies. We evaluate SAF and compare its performance

against baselines IPPO and MAPPO on HECOGrid. Our results show
that SAF consistently outperforms the baselines across different tasks and different
heterogeneity and coordination levels.

****

UNDERSTANDING HTML WITH LARGE LANGUAGE MODELS

Izzeddin Gur,Ofir Nachum,Yingjie Miao,Mustafa Safdari,Austin V Huang,Sharan Narang,Aakanksha Chowdhery,Noah Fiedel,Aleksandra Faust

Large language models (LLM) have shown exceptional performance on a variety of natural language tasks. Yet, their capabilities for HTML understanding – i.e., parsing the raw HTML of a webpage, with applications to automation of web-based tasks, crawling, and browser-assisted retrieval – have not been fully explored. We contribute HTML understanding models (fine-tuned LLMs) and an in-depth analysis of their capabilities under three tasks: (i) Semantic Classification of HTML elements, (ii) Description Generation for HTML inputs, and (iii) Autonomous Web Navigation of HTML pages. While previous work has developed dedicated architectures and training procedures for HTML understanding, we show that LLMs pretrained on standard natural language corpora transfer remarkably well to HTML understanding tasks. For instance, fine-tuned LLMs are 12% more accurate at semantic classification compared to models trained exclusively on the task dataset. Moreover, when fine-tuned on data from the MiniWoB benchmark, LLMs successfully complete 50% more tasks using 192x less data compared to the previous best supervised model. To promote further research on LLMs for HTML understanding, we create and open-source a large-scale HTML dataset distilled and auto-labeled from CommonCrawl. We show evidence that T5-based models due to the bidirectional encoder-decoder architecture are the best choice and that for practitioners larger models are not necessarily better.

****

Kuiper: Moderated Asynchronous Federated Learning on Heterogeneous Mobile Devices with Non-IID Data

Dipesh Tamboli,Pranjal Jain,Atul Sharma,Biplab Banerjee,Saurabh Bagchi,Somali Chaterji

Federated learning allows multiple clients to jointly learn an ML model while keeping their data private. While synchronous federated learning (Sync-FL) requires the devices to share local gradients synchronously to provide better guarantees, it suffers from the problem of stragglers. This is the scenario where the faster clients have to wait for the slower ones, slowing the entire training process. Conventional techniques completely drop the updates from the stragglers and lose the opportunity to learn from the data they hold, which is especially important in a non-iid setting. Asynchronous learning (Async-FL) provides a potential solution to allow the clients to function at their own pace, which typically achieves faster convergence. Since edge devices have a low compute, it is hard to train a video action recognition task on them. We present Kuiper, a variant of Async-FL, to help heterogeneous edge devices with limited resources learn a heavy model on video-action-recognition tasks with data distributed non-IID. Kuiper introduces a novel aggregation scheme, which solves the straggler problem while considering the different data distribution at different clients. Kuiper shows a 11% faster convergence compared to Oort15 [OSDI-21], up to 12% and 9% improvement in test accuracy compared to FedBuff16 [AISTAT-22] and Oort [OSDI-21] on HMDB51, and 10% and 9% on UCF101.

****

Learning Achievement Structure for Structured Exploration in Domains with Sparse Reward

Zihan Zhou,Animesh Garg

We propose Structured Exploration with Achievements (SEA), a multi-stage reinforcement learning algorithm designed for achievement-based environments, a particular type of environment with an internal achievement set. SEA first uses offline data to learn a representation of the known achievements with a determinant loss function, then recovers the dependency graph of the learned achievements with a heuristic algorithm, and finally interacts with the environment online to lear n policies that master known achievements and explore new ones with a controller built with the recovered dependency graph. We empirically demonstrate that SEA can recover the achievement structure accurately and improve exploration in hard domains such as Crafter that are procedurally generated with high-dimensional observations like images.

****

Semi-Autoregressive Energy Flows: Towards Determinant-Free Training of Normalizing Flows

Phillip Si,Zeyi Chen,Subham Sekhar Sahoo,Yair Schiff,Volodymyr Kuleshov

Normalizing flows are a popular approach for constructing probabilistic and generative models. However, maximum likelihood training of flows is challenging due to the need to calculate computationally expensive determinants of Jacobians. This paper takes steps towards addressing this challenge by introducing objectives and model architectures for determinant-free training of flows. Central to our framework is the energy objective, a multidimensional extension of proper scoring rules that admits efficient estimators based on random projections. The energy objective does not require calculating determinants and therefore supports general flow architectures that are not well-suited to maximum likelihood training. In particular, we introduce semi-autoregressive flows, an architecture that can be trained with the energy loss, and that interpolates between fully autoregressive and non-autoregressive models, capturing the benefits of both. We empirically demonstrate that energy flows achieve competitive generative modeling performance while maintaining fast generation and posterior inference.

****

PINTO: Faithful Language Reasoning Using Prompt-Generated Rationales

PeiFeng Wang,Aaron Chan,Filip Ilievski,Muhao Chen,Xiang Ren

Neural language models (LMs) have achieved impressive results on various language-based reasoning tasks by utilizing latent knowledge encoded in their own pretrained parameters. To make this reasoning process more explicit, recent works retrieve a rationalizing LM's internal knowledge by training or prompting it to generate free-text rationales, which can be used to guide task predictions made by either the same LM or a separate reasoning LM. However, rationalizing LMs require expensive rationale annotation and/or computation, without any assurance that their generated rationales improve LM task performance or faithfully reflect LM decision-making. In this paper, we propose PINTO, an LM pipeline that rationalizes via prompt-based learning, and learns to faithfully reason over rationales via a counterfactual regularization. First, PINTO maps out a suitable reasoning process for the task input by prompting a frozen rationalizing LM to generate a free-text rationale. Second, PINTO's reasoning LM is fine-tuned to solve the task using the generated rationale as context, while regularized to output less confident predictions when the rationale is perturbed. Across four datasets, we show that PINTO significantly improves the generalization ability of the reasoning LM, yielding higher performance on both in-distribution and out-of-distribution test sets. Also, we find that PINTO's rationales are more faithful to its task predictions than those generated by competitive baselines.

****

Game Theoretic Mixed Experts for Combinational Adversarial Machine Learning

Ethan Rathbun,Kaleel Mahmood,Sohaib Ahmad,Caiwen Ding,Marten van Dijk

Recent advances in adversarial machine learning have shown that defenses considered to be robust are actually susceptible to adversarial attacks which are specifically tailored to target their weaknesses. These defenses include Barrage of Random Transforms (BaRT), Friendly Adversarial Training (FAT), Trash is Treasure (TiT) and ensemble models made up of Vision Transformers (ViTs), Big Transfer models and Spiking Neural Networks (SNNs). It remains an open question, however, as to whether the adversarial examples designed to target one defense will be similarly misclassified by another defense. In this paper, we provide the first adversarial defense transferability study, as well as a game theoretic framework for ensemble adversarial attacks and defenses. Our framework is called Game theoretic Mixed Experts (GaME) and is designed to find the Mixed-Nash strategy for an attacker that can employ compositional adversarial attacks. We show that this fr

amework creates an ensemble of defenses with greater robustness than a combinational defense with a uniform or random probability distribution. Overall, our framework and analyses advance the field of adversarial machine learning by yielding new insights into compositional attack and defense formulations.
****

Return Augmentation gives Supervised RL Temporal Compositionality
Keiran Paster,Silviu Pitis,Sheila A. McIlraith,Jimmy Ba

Offline Reinforcement Learning (RL) methods that use supervised learning or sequence modeling (e.g., Decision Transformer) work by training a return-conditioned policy. A fundamental limitation of these approaches, as compared to value-based methods, is that they have trouble generalizing to behaviors that have a higher return than what was seen at training. Value-based offline-RL algorithms like CQL use bootstrapping to combine training data from multiple trajectories to learn strong behaviors from sub-optimal data. We set out to endow RL via Supervised Learning (RvS) methods with this form of temporal compositionality. To do this, we introduce SuperB, a dynamic programming algorithm for data augmentation that augments the returns in the offline dataset by combining rewards from intersecting trajectories. We show theoretically that SuperB can improve sample complexity and enable RvS to find optimal policies in cases where it previously fell behind the performance of value-based methods. Empirically, we find that SuperB improves the performance of RvS in several offline RL environments, surpassing the prior state-of-the-art RvS agents in AntMaze by orders of magnitude and offering performance competitive with value-based algorithms on the D4RL-gym tasks.
****

Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods
Shunta Akiyama,Taiji Suzuki

While deep learning has outperformed other methods for various tasks, theoretical frameworks that explain its reason have not been fully established. We investigate the excess risk of two-layer ReLU neural networks in a teacher-student regression model, in which a student network learns an unknown teacher network through its outputs. Especially, we consider the student network that has the same width as the teacher network and is trained in two phases: first by noisy gradient descent and then by the vanilla gradient descent. Our result shows that the student network provably reaches a near-global optimal solution and outperforms any kernel methods estimator (more generally, linear estimators), including neural tangent kernel approach, random feature model, and other kernel methods, in a sense of the minimax optimal rate. The key concept inducing this superiority is the non-convexity of the neural network models. Even though the loss landscape is highly non-convex, the student network adaptively learns the teacher neurons.
****

Automatic Data Augmentation via Invariance-Constrained Learning
Ignacio Hounie,Luiz F. O. Chamon,Alejandro Ribeiro

Underlying data structures, such as symmetries or invariances to transformations, are often exploited to improve the solution of learning tasks. However, embedding these properties in models or learning algorithms can be challenging and computationally intensive. Data augmentation, on the other hand, induces these symmetries during training by applying multiple transformations to the input data. Despite its ubiquity, its effectiveness depends on the choices of which transformations to apply, when to do so, and how often. In fact, there is both empirical and theoretical evidence that the indiscriminate use of data augmentation can introduce biases that outweigh its benefits. This work tackles these issues by automatically adapting the data augmentation while solving the learning task. To do so, it formulates data augmentation as an invariance-constrained learning problem and leverages Monte Carlo Markov Chain (MCMC) sampling to solve it. The result is a practical algorithm that not only does away with a priori searches for augmentation distributions, but also dynamically controls if and when data augmentation is applied. Our experiments illustrate the performance of this method, which achieves state-of-the-art results in automatic data augmentation benchmarks for CIFAR datasets. Furthermore, this approach can be used to gather insights on

the actual symmetries underlying a learning task.
****

GEASS: Neural causal feature selection for high-dimensional biological data
Mingze Dong,Yuval Kluger

Identifying nonlinear causal relationships in high-dimensional biological data is an important task. However, current neural network based causality detection approaches for such data suffer from poor interpretability and cannot scale well to the high dimensional regime. Here we present GEASS (Granger fEAture Selection of Spatiotemporal data), which identifies sparse Granger causality mechanisms of high dimensional spatiotemporal data by a single neural network. GEASS maximizes sparsity-regularized modified transfer entropy with a theoretical guarantee of recovering features with spatial/temporal Granger causal relationships. The sparsity regularization is achieved by a novel combinatorial stochastic gate layer to select sparse non-overlapping feature subsets. We demonstrate the efficacy of GEASS in several synthetic datasets and real biological data from single-cell RNA sequencing and spatial transcriptomics.
****

Unsupervised 3D Scene Representation Learning via Movable Object Inference
Honglin Chen,Wanhee Lee,Hong-Xing Yu,Rahul Mysore Venkatesh,Joshua B. Tenenbaum,
Daniel Bear,Jiajun Wu,Daniel LK Yamins

Unsupervised, category-agnostic, object-centric 3D representation learning for complex scenes remains an open problem in computer vision. While a few recent methods can now discover 3D object radiance fields from a single image without supervision, they are limited to simplistic scenes with objects of a single category, often with a uniform color. This is because they discover objects purely based on appearance cues—objects are made of pixels that look alike. In this work, we propose Movable Object Radiance Fields (MORF), aiming at scaling to complex scenes with diverse categories of objects. Inspired by cognitive science of object learning in babies, MORF learns 3D object representations via movable object inference. During training, MORF first obtains 2D masks of movable objects via a self-supervised movable object segmentation method; it then bridges the gap to 3D object representations via conditional neural rendering in multiple views. During testing, MORF can discover, reconstruct, and move unseen objects from novel categories, all from a single image. Experiments show that MORF extracts accurate object geometry and supports realistic object and scene reconstruction and editing, significantly outperforming the state-of-the-art.
****

Linearly Mapping from Image to Text Space
Jack Merullo,Louis Castricato,Carsten Eickhoff,Ellie Pavlick

The extent to which text-only language models (LMs) learn to represent the physical, non-linguistic world is an open question. Prior work has shown that pretrained LMs can be taught to ``understand'' visual inputs when the models' parameters are updated on image captioning tasks. We test a stronger hypothesis: that the conceptual representations learned by text-only models are functionally equivalent (up to a linear transformation) to those learned by models trained on vision tasks. Specifically, we show that the image representations from vision models can be transferred as continuous prompts to frozen LMs by training only a single linear projection. Using these to prompt the LM achieves competitive performance on captioning and visual question answering tasks compared to models that tune both the image encoder and text decoder (such as the MAGMA model). We compare three image encoders with increasing amounts of linguistic supervision seen during pretraining: BEIT (no linguistic information), NF-ResNET (lexical category information), and CLIP (full natural language descriptions). We find that all three encoders perform equally well at transferring visual property information to the language model (e.g., whether an animal is large or small), but that image encoders pretrained with linguistic supervision more saliently encode category information (e.g., distinguishing hippo vs.\ elephant) and thus perform significantly better on benchmark language-and-vision tasks. Our results indicate that LMs encode conceptual information structurally similarly to vision-based models, even those that are solely trained on images.

****

Actor-Critic Alignment for Offline-to-Online Reinforcement Learning

Zishun Yu,Xinhua Zhang

Deep offline reinforcement learning has recently demonstrated considerable promise in leveraging offline datasets, providing high-quality models that significantly reduce the online interactions required for fine-tuning. However, such a benefit is often diminished due to the marked state-action distribution shift, which causes significant bootstrap error and wipes out the good initial policy. Existing solutions resort to constraining the policy shift or balancing the sample replay based on their online-ness. However, they require online estimation of distribution divergence or density ratio. To avoid such complications, we propose deviating from existing actor-critic approaches that directly transfer the state-action value functions. Instead, we post-process them by aligning with the offline learned policy, so that the Q-values for actions *outside* the offline policy are also tamed. As a result, the online fine-tuning can be simply performed as in the standard actor-critic algorithms. We show empirically that the proposed method improves the performance of the fine-tuned robotic agents on various simulated tasks.

****

Characterizing intrinsic compositionality in transformers with Tree Projections

Shikhar Murty,Pratyusha Sharma,Jacob Andreas,Christopher D Manning

When trained on language data, do transformers learn some arbitrary computation that utilizes the full capacity of the architecture or do they learn a simpler, tree-like computation, hypothesized to underlie compositional meaning systems like human languages? There is an apparent tension between compositional accounts of human language understanding, which are based on a restricted bottom-up computational process, and the enormous success of neural models like transformers, which can route information arbitrarily between different parts of their input. One possibility is that these models, while extremely flexible in principle, in practice learn to interpret language hierarchically, ultimately building sentence representations close to those predictable by a bottom-up, tree-structured model. To evaluate this possibility, we describe an unsupervised and parameter-free method to \emph{functionally project} the behavior of any transformer into the space of tree-structured networks. Given an input sentence, we produce a binary tree that approximates the transformer's representation-building process and a score that captures how ``tree-like'' the transformer's behavior is on the input. While calculation of this score does not require training any additional models, it provably upper-bounds the fit between a transformer and any tree-structured approximation. Using this method, we show that transformers for three different tasks become more tree-like over the course of training, in some cases unsupervisedly recovering the same trees as supervised parsers. These trees, in turn, are predictive of model behavior, with more tree-like models generalizing better on tests of compositional generalization.

****

What Do We Maximize in Self-Supervised Learning And Why Does Generalization Emerge?

Ravid Shwartz-Ziv,Randall Balestriero,Kenji Kawaguchi,Yann LeCun

In this paper, we provide an information-theoretic (IT) understanding of self-supervised learning methods, their construction, and optimality. As a first step, we demonstrate how IT quantities can be obtained for deterministic networks, as an alternative to the commonly used unrealistic stochastic networks assumption. Secondly, we demonstrate how different SSL models can be (re)discovered based on first principles and highlight what the underlying assumptions of different SSL variants are. Third, we derive a novel generalization bound based on our IT understanding of SSL methods, providing generalization guarantees for the downstream supervised learning task. As a result of this bound, along with our unified view of SSL, we can compare the different approaches and provide general guidelines to practitioners. Consequently, our derivation and insights can contribute to a better understanding of SSL and transfer learning from a theoretical and practical perspective.

****

SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing

Sheng Li,Geng Yuan,Yue Dai,Youtao Zhang,Yanzhi Wang,Xulong Tang

There has been a proliferation of artificial intelligence applications, where model training is key to promising high-quality services for these applications. However, the model training process is both time-intensive and energy-intensive, inevitably affecting the user's demand for application efficiency. Layer freezing, an efficient model training technique, has been proposed to improve training efficiency. Although existing layer freezing methods demonstrate the great potential to reduce model training costs, they still remain shortcomings such as lacking generalizability and compromised accuracy. For instance, existing layer freezing methods either require the freeze configurations to be manually defined before training, which does not apply to different networks, or use heuristic freezing criteria that is hard to guarantee decent accuracy in different scenarios. Therefore, there lacks a generic and smart layer freezing method that can automatically perform ``in-situation'' layer freezing for different networks during training processes. To this end, we propose a generic and efficient training framework (SmartFRZ). The core proposed technique in SmartFRZ is attention-guided layer freezing, which can automatically select the appropriate layers to freeze without compromising accuracy. Experimental results show that SmartFRZ effectively reduces the amount of computation in training and achieves significant training acceleration, and outperforms the state-of-the-art layer freezing approaches.

****

Similarity-Based Cooperation

Caspar Oesterheld,Johannes Treutlein,Roger Baker Grosse,Vincent Conitzer,Jakob Nicolaus Foerster

As ML agents act more autonomously in the world, they will increasingly interact with each other. Unfortunately, in many social dilemmas like the one-shot Prisoner’s Dilemma, standard game theory predicts that ML agents will fail to cooperate with each other. Prior work has shown that one way to enable cooperative outcomes in the one-shot Prisoner’s Dilemma is to make the agents mutually transparent to each other, i.e., to allow them to access one another’s source code (Rubinstein 1997; Tennenholtz 2004) or weights in the case of ML agents. However, full transparency is often unrealistic, whereas partial transparency is commonplace. Moreover, it is difficult to machine-learn to cooperate in the full transparency setting. In this paper, we introduce a more realistic setting in which agents only observe a single number indicating how similar they are to each other. We prove that this allows for the same set of cooperative outcomes as the full transparency setting. We also demonstrate experimentally that cooperation can be learned using simple ML methods.

****

Consistent Data Distribution Sampling for Large-scale Retrieval

Hongyu Ou,Jun Yin,Huanqin Wu,ANAN LIU,Lin Zhao,Tao Chen,Yuekui Yang,TAO YANG

Retrieving candidate items with low latency and computational cost is important for large-scale advertising systems. Negative sampling is a general approach to model million-scale items with rich features in the retrieval. The training-inference inconsistency of data distribution brought from sampling negatives is a key challenge. In this work, we propose a novel negative sampling strategy Consistent Data Distribution Sampling (CDDS) to solve such an issue. Specifically, we employ a relative large-scale of uniform training negatives and batch negatives to adequately train long-tail and hot items respectively, and employ high divergence negatives to improve the learning convergence. To make the above training samples approximate the serving item data distribution, we introduce an auxiliary loss based on an asynchronous item embedding matrix over the entire item pool. Offline experiments on real datasets achieve SOTA performance. Online experiments with multiple advertising scenarios show that our method has achieved significant increases in GMV. The source code will be released in the future.

****

NOVEL FEATURE REPRESENTATION STRATEGIES FOR TIME SERIES FORECASTING WITH PREDICTED FUTURE COVARIATES

Jimeng Shi,Rukmangadh Sai Myana,Vitalii Stebliankin,Azam Shirali,Giri Narasimhan

Accurate time series forecasting is a fundamental challenge in data science. Unlike traditional statistical methods, conventional machine learning models, such as RNNs and CNNs, use historical data consisting of previously measured variables including the forecast variable and all its covariates. However, in many applications, some of the covariates can be predicted with reasonable accuracy for the immediate future. Note that the input may also contain some covariates that cannot be accurately predicted. We consider the problem of predicting water levels at a given location in a river or canal system using historical data and future covariates, some of which (e.g., precipitation, tide) may be predictable. In many applications, for some covariates of interest, it may be possible to use historical data or accurate predictions for the near future. Traditional methods to incorporate future predictable covariates have major limitations. The strategy of simply concatenating the future predicted covariates to the input vector is highly likely to miss the past-future connection. Another strategy that iteratively predicts one step at a time can end up with prediction error accumulation. We propose two novel feature representation strategies to solve those limitations -- shifting and padding, which create a framework for contextually linking the past with the predicted future, while avoiding any accumulation of prediction errors. Extensive experiments on three well-known datasets revealed that our strategies when applied to RNN and CNN backbones, outperform existing methods. Our experiments also suggest a relationship between the amount of shifting and padding and the periodicity of the time series.

****

Augmentation Component Analysis: Modeling Similarity via the Augmentation Overlaps

Lu Han,Han-Jia Ye,De-Chuan Zhan

Self-supervised learning aims to learn a embedding space where semantically similar samples are close. Contrastive learning methods pull views of samples together and push different samples away, which utilizes semantic invariance of augmentation but ignores the relationship between samples. To better exploit the power of augmentation, we observe that semantically similar samples are more likely to have similar augmented views. Therefore, we can take the augmented views as a special description of a sample. In this paper, we model such a description as the augmentation distribution, and we call it augmentation feature. The similarity in augmentation feature reflects how much the views of two samples overlap and is related to their semantical similarity. Without computational burdens to explicitly estimate values of the augmentation feature, we propose Augmentation Component Analysis (ACA) with a contrastive-like loss to learn principal components and an on-the-fly projection loss to embed data. ACA equals an efficient dimension reduction by PCA and extracts low-dimensional embeddings, theoretically preserving the similarity of augmentation distribution between samples. Empirical results show that our method can achieve competitive results against various traditional contrastive learning methods on different benchmarks.

****

Replicable Bandits

Hossein Esfandiari,Alkis Kalavasis,Amin Karbasi,Andreas Krause,Vahab Mirrokni,Grigoris Velegkas

In this paper, we introduce the notion of replicable policies in the context of stochastic bandits, one of the canonical problems in interactive learning. A policy in the bandit environment is called replicable if it pulls, with high probability, the exact same sequence of arms in two different and independent executions (i.e., under independent reward realizations). We show that not only do replicable policies exist, but also they achieve almost the same optimal (non-replicable) regret bounds in terms of the time horizon. More specifically, in the stochastic multi-armed bandits setting, we develop a policy with an optimal problem-dependent regret bound whose dependence on the replicability parameter is also optimal. Similarly, for stochastic linear bandits (with finitely and infinitely ma ny arms) we develop replicable policies that achieve the best-known problem-independent regret bounds with an optimal dependency on the replicability parameter. Our results show that even though randomization is crucial for the exploration-exploitation trade-off, an optimal balance can still be achieved while pulling the exact same arms in two different rounds of executions.

****

NEURAL HAMILTONIAN FLOWS IN GRAPH NEURAL NETWORKS

Qiyu Kang,Kai Zhao,Yang Song,Sijie Wang,Wee Peng Tay

Graph neural networks (GNNs) suffer from oversmoothing and oversquashing problems when node features are updated over too many layers. Embedding spaces can also vary significantly for different data types, leading to the need for different GNN model types. In this paper, we model the embedding of a node feature as a Hamiltonian flow over time. As in physics where Hamiltonian flow conserves the energy over time, its induced GNNs enable a more stable feature updating mechanism. Moreover, since the Hamiltonian flows are defined on a general symplectic manifold, this approach allows us to learn the underlying manifold of the graph in training, in contrast to most of the existing literature that assumes a fixed graph embedding manifold. We test Hamiltonian flows of different forms and demonstrate empirically that our approach achieves better node classification accuracy than popular state-of-the-art GNNs.

****

Convergence Analysis of Split Learning on Non-IID Data

Yipeng Li,Xinchen Lyu

Split Learning (SL) is one promising variant of Federated Learning (FL), where the AI model is split and trained at the clients and the server collaboratively. By offloading the computation-intensive portions to the server, SL enables efficient model training on resource-constrained clients. Despite its booming applications, SL still lacks rigorous convergence analysis on non-IID data, which is critical for hyperparameter selection. In this paper, we first prove that SL exhibits an $\mathcal{O}(1/\sqrt{T})$ convergence rate for non-convex objectives on non-IID data, where T is the number of total steps. By comparing the convergence analysis and experimental results, SL can outperform FL in terms of convergence rate (w.r.t. per-client training/communication rounds, and hence, the computation efficiency) and exhibit comparable accuracy to FL on mildly non-IID data. In contrast, FL prevails on highly non-IID data.

****

Principal Trade-off Analysis

Alexander Strang,David Robert SeWell,Alexander Kim,Kevin Alcedo,David Rosenbluth

The focus on equilibrium solutions in games underemphasizes the importance of understanding their overall structure. A different set of tools is needed for learning and representing the general structure of a game. In this paper we illustrate "Principle Trade-off Analysis" (PTA), a decomposition method that embeds games into a low dimensional feature space and argue that the embeddings are more revealing than previously demonstrated. Here, we develop an analogy to Principal Component Analysis (PCA). PTA represents an arbitrary two-player zero-sum game as the weighted sum of pairs of orthogonal 2D feature planes. We show that each of the feature planes represent unique strategic trade-offs (cyclic modes) and truncation of the sequence provides insightful model reduction. We demonstrate the validity of PTA on a pair of games (Blotto, Pokemon). In Blotto, PTA identifies game symmetries, and specifies strategic trade-offs associated with distinct win conditions. These symmetries reveal limitations of PTA unaddressed in previous work. For Pokemon, PTA recovers clusters that naturally correspond to Pokemon types, correctly identifies the designed tradeoff between those types, and discovers a rock-paper-scissor (RPS) cycle in the Pokemon generation type - all absent any specific information except game outcomes.

****

Neural Bregman Divergences for Distance Learning

Fred Lu,Edward Raff,Francis Ferraro

Many metric learning tasks, such as triplet learning, nearest neighbor retrieval, and visualization, are treated primarily as embedding tasks where the ultimate metric is some variant of the Euclidean distance (e.g., cosine or Mahalanobis), and the algorithm must learn to embed points into the pre-chosen space. The study of non-Euclidean geometries is often not explored, which we believe is due to a lack of tools for learning non-Euclidean measures of distance. Recent work has shown that Bregman divergences can be learned from data, opening a promising approach to learning asymmetric distances. We propose a new approach to learning arbitrary Bergman divergences in a differentiable manner via input convex neural networks and show that it overcomes significant limitations of previous works. We also demonstrate that our method more faithfully learns divergences over a set of both new and previously studied tasks, including asymmetric regression, ranking, and clustering. Our tests further extend to known asymmetric, but non-Bregman tasks, where our method still performs competitively despite misspecification, showing the general utility of our approach for asymmetric learning.

****

Offline Reinforcement Learning from Heteroskedastic Data Via Support Constraints

Anikait Singh,Aviral Kumar,quan vuong,Yevgen Chebotar,Sergey Levine

Offline reinforcement learning (RL) learns policies entirely from static datasets, thereby avoiding the challenges associated with online data collection. Practical applications of offline RL will inevitably require learning from datasets where the variability of demonstrated behaviors changes non-uniformly across the state space. For example, at a red light, nearly all human drivers behave similarly by stopping, but when merging onto a highway, some drivers merge quickly, efficiently, and safely, while many hesitate or merge dangerously.

We show that existing popular offline RL methods based on distribution constraints fail to learn from data with such non-uniform change in the variability of demonstrated behaviors, often due to the requirement to stay close to the behavior policy to the same extent across the state space. We demonstrate this failure mode both theoretically and experimentally. Ideally, the learned policy should be free to choose per-state how closely to follow the behavior policy to maximize long-term return, as long as the learned policy stays within the support of the behavior policy. To instantiate this principle, we reweight the data distribution in conservative Q-learning and show that support constraints emerge when doing so. The reweighted distribution is a mixture of the current policy and an additional policy trained to mine poor actions that are likely under the behavior policy. Our method CQL (ReDS) is simple, theoretically motivated, and improves performance across a wide range of offline RL problems in Atari games, navigation, and pixel-based manipulation.

****

Finding Private Bugs: Debugging Implementations of Differentially Private Stochastic Gradient Descent

Congyu Fang,Hengrui Jia,Ali Shahin Shamsabadi,Nicolas Papernot

It is important to learn with privacy-preserving algorithms when training data contains sensitive information. Differential privacy (DP) proposes to bound the worst-case privacy leakage of a training algorithm. However, the analytic nature of these algorithmic guarantees makes it difficult to verify that an implementation of a differentially private learner is correct. Research in the field focuses on empirically approximating the analytic bound, which only assesses whether an implementation provides the guarantee claimed for a particular dataset or not. It is also typically costly. In this paper, we take a first step towards providing a simple and lightweight methodology for practitioners to identify common implementation mistakes without imposing any changes to their scripts. Our approach stems from measuring distances between models outputted by the training algorithm. We demonstrate that our method successfully identifies specific mistakes made in the implementation of DP-SGD, the de facto algorithm for differentially private deep learning. These mistakes include improper gradient computations or noise miscalibration. Both approaches invalidate assumptions that are essential to obtaining a rigorous privacy guarantee.

****

A Computationally Efficient Sparsified Online Newton Method

Fnu Devvrit,Sai Surya Duvvuri,Rohan Anil,Vineet Gupta,Cho-Jui Hsieh,Inderjit S D hillon

Second-order methods have huge potential in improving the convergence of deep neural network (DNN) training, but are prohibitive due to their large memory and compute requirements. Furthermore, computing the matrix inverse or the Newton direction, which is needed in second-order methods, requires high precision computation for stable training as the preconditioner could have a large condition number. This paper provides a first attempt at developing computationally efficient sparse preconditioners for DNN training which can also tolerate low precision computation. Our new Sparsified Online Newton (SONew) algorithm emerges from the novel use of the so-called LogDet matrix divergence measure; we combine it with sparsity constraints to minimize the regret in the online convex optimization framework. Our mathematical analysis allows us to reduce the condition number of our sparse preconditioning matrix, thus improving the stability of training with low precision. We conduct experiments on a feed-forward neural-network autoencoder benchmark, where we compare training loss of optimizers when run for a fixed number of epochs. In the float32 experiments, our methods outperform the best-performing first-order optimizers and perform comparably to Shampoo, a state-of-the-art second-order optimizer. However, our method is even more effective in low-precision, where SONew finishes training considerably faster while performing comparably with Shampoo on training loss.

****

Solving Continual Learning via Problem Decomposition

Gyuhak Kim,Changnan Xiao,Tatsuya Konishi,Zixuan Ke,Bing Liu

This paper is concerned with class incremental learning (CIL) in continual learning (CL). CIL is the popular continual learning paradigm in which a system receives a sequence of tasks with different classes in each task and is expected to learn to predict the class of each test instance without given any task related information for the instance. Although many techniques have been proposed to solve CIL, it remains to be highly challenging due to the difficulty of dealing with catastrophic forgetting (CF). This paper starts from the first principle and proposes a novel method to solve the problem. The definition of CIL reveals that the problem can be decomposed into two probabilities: within-task prediction probability and task-id prediction probability. This paper proposes an effective technique to estimate these two probabilities based on the estimation of feature distributions in the latent space using incremental PCA and Mahalanobis distance. The proposed method does not require a memory buffer to save replay data and it outperforms strong baselines including replay-based methods.

****

Long Term Fairness via Performative Distributionally Robust Optimization

Garnet Liam Peet-Pare,Nidhi Hegde,Alona Fyshe

Fairness researchers in machine learning (ML) have coalesced around several fairness criteria which provide formal definitions of what it means for an ML model to be fair. However, these criteria have some serious limitations. We identify four key shortcomings of these formal fairness criteria and address them by extending performative prediction to include a distributionally robust objective. Performative prediction is a recent framework developed to understand the effects of when deploying model influences the distribution on which it is making predictions. We prove a convergence result for our proposed repeated distributionally robust optimization (RDRO). We further verify our results empirically and develop experiments to demonstrate the impact of using RDRO on learning fair ML models.

****

The In-Sample Softmax for Offline Reinforcement Learning

Chenjun Xiao,Han Wang,Yangchen Pan,Adam White,Martha White

Reinforcement learning (RL) agents can leverage batches of previously collected data to extract a reasonable control policy. An emerging issue in this offline RL setting, however, is that the bootstrapping update underlying many of our methods suffers from insufficient action-coverage: standard max operator may select a maximal action that has not been seen in the dataset. Bootstrapping from these inaccurate values can lead to overestimation and even divergence. There are a g rowing number of methods that attempt to approximate an in-sample max, that only uses actions well-covered by the dataset. We highlight a simple fact: it is more straightforward to approximate an in-sample softmax using only actions in the dataset. We show that policy iteration based on the in-sample softmax converges, and that for decreasing temperatures it approaches the in-sample max. We derive an In-Sample Actor-Critic (AC), using this in-sample softmax, and show that it is consistently better or comparable to existing offline RL methods, and is also well-suited to fine-tuning. We release the code at github.com/hwang-ua/inac_pytorch.

****

LUNA: Language as Continuing Anchors for Referring Expression Comprehension

Yaoyuan Liang,Zhao Yang,Yansong Tang,Jiashuo Fan,Jingang Wang,Wei Wu,Philip Torr,Shao-Lun Huang

Referring expression comprehension aims to localize the description of a natural language expression in an image. Using location priors to remedy inaccuracies in cross-modal alignments is the state of the art for CNN-based methods tackling this problem. Recent Transformer-based models cast aside this idea making the case for steering away from hand-designed components. In this work, we propose LUNA, which uses language as continuing anchors to guide box prediction in a Transformer decoder, and show that language-guided location priors can be effectively exploited in a Transformer-based architecture. Specifically, we first initialize an anchor box from the input expression via a small “proto-decoder”, and then use this anchor as location prior in a modified Transformer decoder for predicting the bounding box. Iterating through each decoder layer, the anchor box is first used as a query for pooling multi-modal context, and then updated based on pooled context. This approach allows the decoder to focus selectively on one part of the scene at a time, which reduces noise in multi-modal context and leads to more accurate box predictions. Our method outperforms existing state-of-the-art methods on the challenging datasets of ReferIt Game, RefCOCO/+/g, and Flickr30K Entities.

****

Bias Propagation in Federated Learning

Hongyan Chang,Reza Shokri

We show that participating in federated learning can be detrimental to group fairness. In fact, the bias of a few parties against under-represented groups (identified by sensitive attributes such as gender or race) can propagate through the network to all the parties in the network. We analyze and explain bias propagation in federated learning on naturally partitioned real-world datasets. Our analysis reveals that biased parties unintentionally yet stealthily encode their bias in a small number of model parameters, and throughout the training, they steadily increase the dependence of the global model on sensitive attributes. What is important to highlight is that the experienced bias in federated learning is higher than what parties would otherwise encounter in centralized training with a model trained on the union of all their data. This indicates that the bias is due to the algorithm. Our work calls for auditing group fairness in federated learning and designing learning algorithms that are robust to bias propagation.

****

Causal Confusion and Reward Misidentification in Preference-Based Reward Learning

Jeremy Tien,Jerry Zhi-Yang He,Zackory Erickson,Anca Dragan,Daniel S. Brown

Learning policies via preference-based reward learning is an increasingly popular method for customizing agent behavior, but has been shown anecdotally to be prone to spurious correlations and reward hacking behaviors. While much prior work focuses on causal confusion in reinforcement learning and behavioral cloning, we focus on a systematic study of causal confusion and reward misidentification when learning from preferences. In particular, we perform a series of sensitivity and ablation analyses on several benchmark domains where rewards learned from preferences achieve minimal test error but fail to generalize to out-of-distribution states---resulting in poor policy performance when optimized. We find that t he presence of non-causal distractor features, noise in the stated preferences, and partial state observability can all exacerbate reward misidentification. We also identify a set of methods with which to interpret misidentified learned rewards. In general, we observe that optimizing misidentified rewards drives the policy off the reward's training distribution, resulting in high predicted (learned) rewards but low true rewards. These findings illuminate the susceptibility of preference learning to reward misidentification and causal confusion---failure to consider even one of many factors can result in unexpected, undesirable behavior.

****

UniKGQA: Unified Retrieval and Reasoning for Solving Multi-hop Question Answering Over Knowledge Graph

Jinhao Jiang,Kun Zhou,Xin Zhao,Ji-Rong Wen

Multi-hop Question Answering over Knowledge Graph~(KGQA) aims to find the answer entities that are multiple hops away from the topic entities mentioned in a natural language question on a large-scale Knowledge Graph (KG).

To cope with the vast search space, existing work usually adopts a two-stage approach: it first retrieves a relatively small subgraph related to the question and then performs the reasoning on the subgraph to find the answer entities accurately.

Although these two stages are highly related, previous work employs very different technical solutions for developing the retrieval and reasoning models, neglecting their relatedness in task essence.

In this paper, we propose UniKGQA, a novel approach for multi-hop KGQA task, by unifying retrieval and reasoning in both model architecture and parameter learning.

For model architecture, UniKGQA consists of a semantic matching module based on a pre-trained language model~(PLM) for question-relation semantic matching, and a matching information propagation module to propagate the matching information along the directed edges on KGs.

For parameter learning, we design a shared pre-training task based on question-relation matching for both retrieval and reasoning models, and then propose retrieval- and reasoning-oriented fine-tuning strategies.

Compared with previous studies, our approach is more unified, tightly relating the retrieval and reasoning stages.

Extensive experiments on three benchmark datasets have demonstrated the effectiveness of our method on the multi-hop KGQA task.

Our codes and data are publicly available at~\url{https://github.com/RUCAIBox/UniKGQA}.

****

Comparing Human and Machine Bias in Face Recognition

Samuel Dooley,George Zhihong Wei,Ryan Downing,Nathan Shankar,Bradon Michael Thymes,Gudrun Lilja Thorkelsdottir,Tiye A Kurtz-Miott,Rachel Mattson,Olufemi Obiwumi,Valeriia Cherepanova,Micah Goldblum,John P Dickerson,Tom Goldstein

Much recent research has uncovered and discussed serious concerns of bias in facial analysis technologies, finding performance disparities between groups of people based on perceived gender, skin type, lighting condition, etc. These audits are immensely important and successful at measuring algorithmic bias but have two major challenges: the audits (1) use facial recognition datasets which lack quality metadata, like LFW and CelebA, and (2) do not compare their observed algorithmic bias to the biases of their human alternatives. In this paper, we release improvements to the LFW and CelebA datasets which will enable future researchers to obtain measurements of algorithmic bias that are not tainted by major flaws in the dataset (e.g. identical images appearing in both the gallery and test set). We also use these new data to develop a series of challenging facial identification and verification questions that we administered to various algorithms and a large, balanced sample of human reviewers. We find that both computer models and human survey participants perform significantly better at the verification task, generally obtain lower accuracy rates on dark-skinned or female subjects for both tasks, and obtain higher accuracy rates when their demographics match th at of the question. Academic models exhibit comparable levels of gender bias to humans, but are significantly more biased against darker skin types than humans.

****

Sufficient Subgraph Embedding Memory for Continual Graph Representation Learning

Xikun ZHANG,Dongjin Song,Dacheng Tao

Memory replay, which constructs a buffer to store representative samples and retrain the model over the buffer to maintain its performance over existing tasks, has shown great success for continual learning with Euclidean data. Directly applying it to graph data, however, can lead to the memory explosion problem due to the necessity to consider explicit topological connections of representative nodes. To this end, we present Parameter Decoupled Graph Neural Networks (PDGNNs) with Sufficient Subgraph Embedding Memory (SSEM) to fully utilize the explicit topological information for memory replay and reduce the memory space complexity from $\mathcal{O}(nd^L)$ to $\mathcal{O}(n)$, where n is the memory buffer size, d is the average node degree, and L is the range of neighborhood aggregation. Specifically, PDGNNs decouple trainable parameters from the computation subgraphs via $\textit{Sufficient Subgraph Embeddings}$ (SSEs), which compress subgraphs into vectors ($\textit{i.e.}$, SSEs) to reduce the memory consumption. Besides, we discover a $\textit{pseudo-training effect}$ in memory based continual graph learning, which does not exist in continual learning on Euclidean data without topological connection ($\textit{e.g.}$, individual images). Based on the discovery, we develop a novel $\textit{coverage maximization sampling}$ strategy to enhance the performance when the memory budget is tight. Thorough empirical studies demonstrate that PDGNNs with SSEM outperform state-of-the-art techniques for both class-incremental and task-incremental settings.

****

One cannot stand for everyone! Leveraging Multiple User Simulators to train Task-oriented Dialogue Systems

Yajiao LIU,Xin Jiang,Yichun Yin,Yasheng Wang,Fei Mi,Qun Liu,Xiang Wan,Benyou Wang

User simulators are agents designed to imitate human users; recent advances have found that Task-oriented Dialogue (ToD) systems optimized toward a user simulator could better satisfy the need of human users. However, this might result in a sub-optimal ToD system if it is tailored to only one \textit{ad hoc} user simulator, since human users can behave differently.

In this paper, we propose a framework called MUST to optimize ToD systems via leveraging \textbf{m}ultiple \textbf{u}ser \textbf{s}imula\textbf{t}ors.

The main challenges of MUST fall in 1) how to adaptively specify which user simulator to interact with the ToD system at each optimization step, since the ToD system might be over-fitted to some specific user simulators, and simultaneously under-fitted to some others; 2) how to avoid catastrophic forgetting of the adaption for a simulator that is not selected for several consecutive optimization steps.

To tackle these challenges, we formulate MUST as a Multi-armed bandits (MAB) problem and provide a method called MUST$_{\mathrm{adaptive}}$ that balances \textit{i}) the \textit{boosting adaption} for adaptive interactions between different user simulators and the ToD system and \textit{ii}) the \textit{uniform adaption} to avoid the catastrophic forgetting issue.

With both automatic evaluations and human evaluations, our extensive experimental results on the restaurant search task from MultiWOZ show that the dialogue system trained by our proposed MUST achieves a better performance than those trained by any single user simulator. It also has a better generalization ability when testing with unseen user simulators. Moreover, our method MUST$_{\mathrm{adaptive}}$ is indeed more efficient and effective to leverage multiple user simulators by our visualization analysis.

****

Is the Performance of My Deep Network Too Good to Be True? A Direct Approach to Estimating the Bayes Error in Binary Classification

Takashi Ishida,Ikko Yamane,Nontawat Charoenphakdee,Gang Niu,Masashi Sugiyama

There is a fundamental limitation in the prediction performance that a machine learning model can achieve due to the inevitable uncertainty of the prediction target. In classification problems, this can be characterized by the Bayes error, which is the best achievable error with any classifier. The Bayes error can be used as a criterion to evaluate classifiers with state-of-the-art performance and can be used to detect test set overfitting. We propose a simple and direct Bayes error estimator, where we just take the mean of the labels that show \emph{uncertainty} of the class assignments. Our flexible approach enables us to perform Bayes error estimation even for weakly supervised data. In contrast to others, our method is model-free and even instance-free. Moreover, it has no hyperparameters and gives a more accurate estimate of the Bayes error than several baselines empirically. Experiments using our method suggest that recently proposed deep networks such as the Vision Transformer may have reached, or is about to reach, the Bayes error for benchmark datasets. Finally, we discuss how we can study the inherent difficulty of the acceptance/rejection decision for scientific articles, by estimating the Bayes error of the ICLR papers from 2017 to 2023.

****

Learning Deep Operator Networks: The Benefits of Over-Parameterization

Bhavesh Shrimali,Arindam Banerjee

Neural Operators that directly learn mappings between function spaces have received considerable recent attention. Deep Operator Networks (DeepONets), a popular recent class of neural operators have shown promising preliminary results in approximating solution operators of parametric differential equations. Despite the universal approximation guarantees, there is yet no optimization convergence guarantee for DeepONets based on gradient descent (GD). In this paper, we establish such guarantees and show that over-parameterization based on wide layers provably helps. In particular, we present two types of optimization convergence analysis: first, for smooth activations, we bound the spectral norm of the Hessian of DeepONets and use the bound to show geometric convergence of GD based on restricted strong convexity (RSC); and second, for ReLU activations, we show the neural tangent kernel (NTK) of DeepONets at initialization is positive definite, which can be used with the standard NTK analysis to imply geometric convergence. Further, we present empirical results on three canonical operator learning problems: Antiderivative, Diffusion-Reaction equation, and Burger’s equation, and show that wider DeepONets lead to lower training loss on all the problems, thereby supporting the theoretical results

****

How Useful are Gradients for OOD Detection Really?

Conor Igoe,Youngseog Chung,Ian Char,Jeff Schneider

One critical challenge in deploying machine learning models in real-life applications is out-of-distribution (OOD) detection. Given a predictive model which is accurate on in distribution (ID) data, an OOD detection system can further equip the model with the option to defer prediction when the input is novel and the model has low confidence. Notably, there has been some recent interest in utilizing gradient information in pretrained models for OOD detection. While these methods are competitive, we argue that previous works conflate their performance with the necessity of gradients. In this work, we provide an in-depth analysis of gradient-based methods and elucidate the key components that warrant their OOD detection performance. We further demonstrate that a general, non-gradient-based family of OOD detection methods are just as competitive, casting doubt on the usefulness of gradients for OOD detection

****

Many-Body Approximation for Tensors

Kazu Ghalamkari,Mahito Sugiyama

We propose a nonnegative tensor decomposition with focusing on the relationship between the modes of tensors. Traditional decomposition methods assume low-rankness in the representation, resulting in difficulties in global optimization and target rank selection. To address these problems, we present an alternative way to decompose tensors, a many-body approximation for tensors, based on an informa tion geometric formulation. A tensor is treated via an energy-based model, where the tensor and its mode correspond to a probability distribution and a random variable, respectively, and many-body approximation is performed on it by taking the interaction between variables into account. Our model can be globally optimized in polynomial time in terms of the KL divergence minimization, which is empirically faster than low-rank approximations keeping comparable reconstruction error. Furthermore, we visualize interactions between modes as tensor networks and reveal a nontrivial relationship between many-body approximation and low-rank approximation.

****

Faster Last-iterate Convergence of Policy Optimization in Zero-Sum Markov Games

Shicong Cen,Yuejie Chi,Simon Shaolei Du,Lin Xiao

Multi-Agent Reinforcement Learning (MARL)---where multiple agents learn to interact in a shared dynamic environment---permeates across a wide range of critical applications. While there has been substantial progress on understanding the global convergence of policy optimization methods in single-agent RL, designing and analysis of efficient policy optimization algorithms in the MARL setting present significant challenges and new desiderata, which unfortunately, remain highly inadequately addressed by existing theory. In this paper, we focus on the most basic setting of competitive multi-agent RL, namely two-player zero-sum Markov games, and study equilibrium finding algorithms in both the infinite-horizon discounted setting and the finite-horizon episodic setting. We propose a single-loop policy optimization method with symmetric updates from both agents, where the policy is updated via the entropy-regularized optimistic multiplicative weights update (OMWU) method and the value is updated on a slower timescale. We show that, in the full-information tabular setting, the proposed method achieves a finite-time last-iterate linear convergence to the quantal response equilibrium of the regularized problem, which translates to a sublinear convergence to the Nash equilibrium by controlling the amount of regularization. Our convergence results improve upon the best known iteration complexities, and lead to a better understanding of policy optimization in competitive Markov games.

****

Memorization Capacity of Neural Networks with Conditional Computation

Erdem Koyuncu

Many empirical studies have demonstrated the performance benefits of conditional computation in neural networks, including reduced inference time and power consumption. We study the fundamental limits of neural conditional computation from the perspective of memorization capacity. For Rectified Linear Unit (ReLU) networks without conditional computation, it is known that memorizing a collection of n input-output relationships can be accomplished via a neural network with $O(\sqrt{n})$ neurons. Calculating the output of this neural network can be accomplished using $O(\sqrt{n})$ elementary arithmetic operations of additions, multiplications and comparisons for each input. Using a conditional ReLU network, we show that the same task can be accomplished using only $O(\log n)$ operations per input. This represents an almost exponential improvement as compared to networks without conditional computation. We also show that the $\Theta(\log n)$ rate is the best possible. Our achievability result utilizes a general methodology to synthesize a conditional network out of an unconditional network in a computationally-efficient manner, bridging the gap between unconditional and conditional architectures.

****

On the Power of Pre-training for Generalization in RL: Provable Benefits and Hardness

Haotian Ye,Xiaoyu Chen,Liwei Wang,Simon Shaolei Du

Generalization in Reinforcement Learning (RL) aims to train an agent during training that generalizes to the target environment. In this work, we first point out that RL generalization is fundamentally different from the generalization in supervised learning, and fine-tuning on the target environment is necessary for good test performance. Therefore, we seek to answer the following question: how much can we expect pre-training over training environments to be helpful for effi cient and effective fine-tuning? On one hand, we give a surprising result showing that asymptotically, the improvement from pre-training is at most a constant factor. On the other hand, we show that pre-training can be indeed helpful in the non-asymptotic regime by designing a policy collection-elimination (PCE) algorithm and proving a distribution-dependent regret bound that is independent of the state-action space. We hope our theoretical results can provide insight towards understanding pre-training and generalization in RL.

****

A Fast, Well-Founded Approximation to the Empirical Neural Tangent Kernel

Mohamad Amin Mohamadi,Danica J. Sutherland

Empirical neural tangent kernels (eNTKs) can provide a good understanding of a given network's representation: they are often far less expensive to compute and applicable more broadly than infinite-width NTKs. For networks with O output units (e.g. an O-class classifier), however, the eNTK on N inputs is of size $NO \times NO$, taking $\mathcal{O}\big((N O)^2\big)$ memory and up to $\mathcal{O}\big((N O)^3 \big)$ computation. Most existing applications have therefore used one of a handful of approximations yielding $N \times N$ kernel matrices, saving orders of magnitude of computation, but with limited to no justification. We prove that one such approximation, which we call ``sum of logits,'' converges to the true eNTK at initialization. Our experiments demonstrate the quality of this approximation for various uses across a range of settings.

****

Weighted Clock Logic Point Process

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

Datasets involving multivariate event streams are prevalent in numerous applications. We present a novel framework for modeling temporal point processes called clock logic neural networks (CLNN) which learn weighted clock logic (wCL) formulas as interpretable temporal rules by which some events promote or inhibit other events. Specifically, CLNN models temporal relations between events using conditional intensity rates informed by a set of wCL formulas, which are more expressive than related prior work. Unlike conventional approaches of searching for generative rules through expensive combinatorial optimization, we design smooth activation functions for components of wCL formulas that enable a continuous relaxation of the discrete search space and efficient learning of wCL formulas using gradient-based methods. Experiments on synthetic datasets manifest our model's ability to recover the ground-truth rules and improve computational efficiency. In addition, experiments on real-world datasets show that our models perform competitively when compared with state-of-the-art models.

****

Simple Emergent Action Representations from Multi-Task Policy Training

Pu Hua,Yubei Chen,Huazhe Xu

The low-level sensory and motor signals in deep reinforcement learning, which exist in high-dimensional spaces such as image observations or motor torques, are inherently challenging to understand or utilize directly for downstream tasks. While sensory representations have been extensively studied, the representations of motor actions are still an area of active exploration. Our work reveals that a space containing meaningful action representations emerges when a multi-task policy network takes as inputs both states and task embeddings. Moderate constraints are added to improve its representation ability. Therefore, interpolated or composed embeddings can function as a high-level interface within this space, providing instructions to the agent for executing meaningful action sequences. Empirical results demonstrate that the proposed action representations are effective for intra-action interpolation and inter-action composition with limited or no additional learning. Furthermore, our approach exhibits superior task adaptation ability compared to strong baselines in Mujoco locomotion tasks. Our work sheds light on the promising direction of learning action representations for efficient, adaptable, and composable RL, forming the basis of abstract action planning and the understanding of motor signal space. Project page: https://sites.google.com/view/emergent-action-representation/

****

Iterative Task-adaptive Pretraining for Unsupervised Word Alignment

Zhi Li,Gao Xing,Ji Zhang,Yin Zhang

How to establish a closer relationship between pre-training and downstream task is a valuable question. We argue that task-adaptive pretraining should not just performed before task. For word alignment task, we propose an iterative self-supervised task-adaptive pretraining paradigm, tying together word alignment and self-supervised pretraining by code-switching data augmentation. When we get the aligned pairs predicted by the multilingual contextualized word embeddings, we employ these pairs and origin parallel sentences to synthesize code-switched sentences. Then multilingual models will be continuously finetuned on the augmented code-switched dataset. Finally, finetuned models will be used to produce new aligned pairs. This process will be executed iteratively. Our paradigm is suitable for almost all unsupervised word alignment methods based on multilingual pre-trained LMs and doesn't need gold labeled data, extra parallel data or any other external resources. Experimental results on six language pairs demonstrate that our paradigm can consistently improve baseline method. Compared to resource-rich languages, the improvements on relatively low-resource or different morphological languages are more significant. For example, the AER scores of three different alignment methods based on XLM-R are reduced by about $4 \sim 5$ percentage points on language pair En-Hi.

****

Open-Set 3D Detection via Image-level Class and Debiased Cross-modal Contrastive Learning

Yuheng Lu,Chenfeng Xu,Xiaobao Wei,Xiaodong Xie,Masayoshi Tomizuka,Kurt Keutzer,Shanghang Zhang

Current point-cloud detection methods have difficulty detecting the open-set objects in the real world, due to their limited generalization capability. Moreover, it is extremely laborious and expensive to collect and fully annotate a point-cloud detection dataset with numerous classes of objects, leading to the limited classes of existing point-cloud datasets and hindering the model to learn general representations to achieve open-set point-cloud detection. Instead of seeking a point-cloud dataset with full labels, we resort to ImageNet1K to broaden the vocabulary of the point-cloud detector. We propose OS-3DETIC, an Open-Set 3D DETector using Image-level Class supervision. Specifically, we take advantage of two modalities, the image modality for recognition and the point-cloud modality for localization, to generate pseudo labels for unseen classes. Then we propose a novel debiased cross-modal cross-task contrastive learning method to transfer the knowledge from image modality to point-cloud modality during training. Without hurting the latency during inference, OS-3DETIC makes the well-known point-cloud detector capable of achieving open-set detection. Extensive experiments demonstrate that the proposed OS-3DETIC achieves at least 10.77 % mAP improvement (absolute value) and 9.56 % mAP improvement (absolute value) by a wide range of baselines on the SUN-RGBD dataset and ScanNet dataset, respectively. Besides, we conduct sufficient experiments to shed light on why the proposed OS-3DETIC works.

****

Tight Non-asymptotic Inference via Sub-Gaussian Intrinsic Moment Norm

Huiming Zhang,Haoyu Wei,Guang Cheng

In non-asymptotic statistical inferences, variance-type parameters of sub-Gaussian distributions play a crucial role. However, direct estimation of these parameters based on the empirical moment generating function (MGF) is infeasible. To this end, we recommend using a sub-Gaussian intrinsic moment norm [Buldygin and Kozachenko (2000), Theorem 1.3] through maximizing a series of normalized moments. Importantly, the recommended norm can not only recover the exponential moment bounds for the corresponding MGFs, but also lead to tighter Hoeffiding's sub-Gaussian concentration inequalities. In practice, intrinsic moment norm can be robustly and consistently estimated via a simple plug-in approach. Our theoretical results are applied to non-asymptotic analysis, including the multi-armed bandit.

****

Interaction-Based Disentanglement of Entities for Object-Centric World Models

Akihiro Nakano,Masahiro Suzuki,Yutaka Matsuo

Perceiving the world compositionally in terms of space and time is essential to understanding object dynamics and solving downstream tasks. Object-centric learning using generative models has improved in its ability to learn distinct representations of individual objects and predict their interactions, and how to utilize the learned representations to solve untrained, downstream tasks is a focal question. However, as models struggle to predict object interactions and track the objects accurately, especially for unseen configurations, using object-centric representations in downstream tasks is still a challenge. This paper proposes STEDIE, a new model that disentangles object representations, based on interactions, into interaction-relevant relational features and interaction-irrelevant global features without supervision. Empirical evaluation shows that the proposed model factorizes global features, unaffected by interactions from relational features that are necessary to predict outcome of interactions. We also show that STEDIE achieves better performance in planning tasks and understanding causal relationships. In both tasks, our model not only achieves better performance in terms of reconstruction ability but also utilizes the disentangled representations to solve the tasks in a structured manner.

****

CodeT5Mix: A Pretrained Mixture of Encoder-decoder Transformers for Code Understanding and Generation

Yue Wang,Hung Le,Akhilesh Deepak Gotmare,Junnan Li,Steven Hoi

Pretrained language models (LMs) trained on vast source code have achieved prominent progress in a wide range of code intelligence tasks. Despite their success, they either adopt specific types of network architectures (encoder-only or decoder-only) for different downstream tasks or rely on a single architecture (encoder-decoder or UniLM-style encoder) for all tasks. The latter approach usually results in a sub-optimal performance on a subset of tasks. To address these limitations, we propose “CodeT5Mix”, a mixture of encoder-decoder Transformers for code where its components can be flexibly combined based on the target tasks during finetuning, while still enjoying the mutual benefits from the joint pretraining. To endow the model with both code understanding and generation capabilities, we pretrain CodeT5Mix using a mixture of denoising, contrastive learning, matching, and Causal Language Modeling (CLM) tasks on large-scale multilingual code corpora in nine programming languages. Additionally, we design a weight sharing strategy in decoders except the feedforward layers, which act as task-specific experts to reduce the interference across tasks of various types. We extensively evaluate CodeT5Mix on seven tasks in four different modes and achieve state-of-the-art (SoTA) performance on most tasks such as text-to-code retrieval, code completion and generation, and math programming. Particularly, we demonstrate that CodeT5Mix can be used as a unified semi-parametric retrieval-augmented generator with SoTA code generation performance.

****

Neural Image-based Avatars: Generalizable Radiance Fields for Human Avatar Modeling

YoungJoong Kwon,Dahun Kim,Duygu Ceylan,Henry Fuchs

We present a method that enables synthesizing novel views and novel poses of arbitrary human performers from sparse multi-view images. A key ingredient of our method is a hybrid appearance blending module that combines the advantages of the implicit body NeRF representation and image-based rendering. Existing generalizable human NeRF methods that are conditioned on the body model have shown robustness against the geometric variation of arbitrary human performers. Yet they often exhibit blurry results when generalized onto unseen identities. Meanwhile, image-based rendering shows high-quality results when sufficient observations are available, whereas it suffers artifacts in sparse-view settings. We propose Neural Image-based Avatars (NIA) that exploits the best of those two methods: to maintain robustness under new articulations and self-occlusions while directly leveraging the available (sparse) source view colors to preserve appearance details of new subject identities. Our hybrid design outperforms recent methods on both in-domain identity generalization as well as challenging cross-dataset generalization settings. Also, in terms of the pose generalization, our method outperforms even the per-subject optimized animatable NeRF methods.

****

Federated Neural Bandits

Zhongxiang Dai,Yao Shu,Arun Verma,Flint Xiaofeng Fan,Bryan Kian Hsiang Low,Patrick Jaillet

Recent works on neural contextual bandits have achieved compelling performances due to their ability to leverage the strong representation power of neural networks (NNs) for reward prediction. Many applications of contextual bandits involve multiple agents who collaborate without sharing raw observations, thus giving rise to the setting of federated contextual bandits}. Existing works on federated contextual bandits rely on linear or kernelized bandits, which may fall short when modeling complex real-world reward functions. So, this paper introduces the federated neural-upper confidence bound (FN-UCB) algorithm. To better exploit the federated setting, FN-UCB adopts a weighted combination of two UCBs: UCB^{a} allows every agent to additionally use the observations from the other agents to accelerate exploration (without sharing raw observations), while UCB^{b} uses an NN with aggregated parameters for reward prediction in a similar way to federated averaging for supervised learning. Notably, the weight between the two UCBs required by our theoretical analysis is amenable to an interesting interpretation, which emphasizes UCB^{a} initially for accelerated exploration and relies more on UCB^{b} later after enough observations have been collected to train the NNs for accurate reward prediction (i.e., reliable exploitation). We prove sub-linear upper bounds on both the cumulative regret and the number of communication rounds of FN-UCB, and empirically demonstrate its competitive performance.

****

Compositional Task Representations for Large Language Models

NAN SHAO,Zefan Cai,Hanwei xu,Chonghua Liao,Yanan Zheng,Zhilin Yang

Large language models have shown a remarkable cross-task generalization ability. Most prior work assumed that prompts effectively extract knowledge from language models to facilitate generalization to new tasks. This perspective led to numerous studies on improving prompts. In contrast, we introduce a new perspective, compositional generalization, that views each task as a composition of latent codes and generalizes to test tasks by a new composition of seen codes. To this end, we propose a novel prompt-free approach, Compositional Task Representations (CTR), that employs multi-task training to learn a discrete, compositional codebook. Empirically, our CTR substantially outperforms prompt-based methods in zero-label learning on average. According to our analysis, some of the learned CTR codes are interpretable to human and demonstrate a certain degree of controllability.

****

What do large networks memorize?

Michal Lukasik,Aditya Krishna Menon,Ankit Singh Rawat,Vaishnavh Nagarajan,Sanjiv Kumar

The success of modern neural models has prompted renewed study of the connection between memorisation and generalisation: such models typically generalise well, despite being able to perfectly fit ("memorise") completely random labels.

To more carefully study this issue, Feldman (2019); Feldman & Zhang (2020) provided a simple metric to quantify the degree of memorisation of a specific training example, and empirically quantified the corresponding memorisation profile of a ResNet model on image classification benchmarks.

While an exciting first glimpse into how real-world models memorise, these studies leave open several questions about memorisation of practical networks.

In particular, how is memorisation affected by increasing model size, and by distilling a large model into a smaller one?

We present a systematic empirical analysis of these questions.

On standard image classification benchmarks, we find that training examples exhi bit a diverse set of memorisation trajectories across model sizes, with some samples having increased memorisation under larger models.

Further, we find that distillation tends to inhibit memorisation of the student model, while also improving generalisation.

Finally, we show that computationally tractable measures of memorisation do not capture the properties we identify for memorisation in the sense of Feldman (2019), despite highly correlating to the latter.

****

TILDE-Q: a Transformation Invariant Loss Function for Time-Series Forecasting

Hyunwook Lee,Chunggi Lee,Hongkyu Lim,Sungahn Ko

Time-series forecasting has caught increasing attention in the AI research field due to its importance in solving real-world problems across different domains, such as energy, weather, traffic, and economy. As shown in various types of data, it has been a must-see issue to deal with drastic changes, temporal patterns, and shapes in sequential data that previous models are weak in prediction. This is because most cases in time-series forecasting aim to minimize L_p norm distances as loss functions, such as mean absolute error (MAE) or mean square error (MSE). These loss functions are vulnerable to not only considering temporal dynamics modeling but also capturing the shape of signals. In addition, these functions often make models misbehave and return uncorrelated results to the original time-series. To become an effective loss function, it has to be invariant to the set of distortions between two time-series data instead of just comparing exact values. In this paper, we propose a novel loss function, called TILDE-Q (Transformation Invariant Loss function with Distance EQuilibrium), that not only considers the distortions in amplitude and phase but also allows models to capture the shape of time-series sequences. In addition, TILDE-Q supports modeling periodic and non-periodic temporal dynamics at the same time. We evaluate the effectiveness of TILDE-Q by conducting extensive experiments with respect to periodic and non-periodic conditions of data, from naive models to state-of-the-art models. The experiment results indicate that the models trained with TILDE-Q outperform those trained with other training metrics (e.g., MSE, dynamic time warping (DTW), temporal distortion index (TDI), and longest common subsequence (LCSS)).

****

A Picture of the Space of Typical Learning Tasks

Rahul Ramesh,Jialin Mao,Itay Griniasty,Rubing Yang,Han Kheng Teoh,Mark Transtrum,James Sethna,Pratik Chaudhari

We develop a technique to analyze representations learned by deep networks when they are trained on different tasks using supervised, multi-task, meta- and contrastive learning. We develop a technique to visualize such representations using an isometric embedding of the space of probabilistic models into a lower-dimensional space, i.e., one that preserves pairwise distances. We discover the following surprising phenomena that shed light upon the structure in the space of learning tasks: (1) the manifold of probabilistic models trained on different tasks using different representation learning methods is effectively low-dimensional; (2) supervised learning on one task results in a surprising amount of progress on seemingly dissimilar tasks; progress on other tasks is larger if the training task has diverse classes; (3) the structure of the space of tasks indicated by our visualization technique is consistent with parts of the Wordnet phylogenetic tree; (4) fine-tuning a model upon a sub-task does not change the representation much if the model was trained for a large number of epochs; (5) episodic meta-learning algorithms fit similar models eventually as that of supervised learning, even if the two traverse different trajectories during training; (6) contrastive learning methods trained on different datasets learn similar representations. We use classification tasks constructed from the CIFAR-10 and Imagenet datasets to study these phenomena.

****

REPAIR: REnormalizing Permuted Activations for Interpolation Repair

Keller Jordan,Hanie Sedghi,Olga Saukh,Rahim Entezari,Behnam Neyshabur

In this paper we empirically investigate the conjecture from Entezari et al. (2021) which states that if permutation invariance is taken into account, then ther e should be no loss barrier to the linear interpolation between SGD solutions. We conduct our investigation using standard computer vision architectures trained on CIFAR-10 and ImageNet. First, we observe a general phenomenon in which interpolated deep networks suffer a collapse in the variance of their activations. We demonstrate that an appropriate rescaling of the pre-activations of the interpolated networks ameliorates this problem and significantly reduces the barrier. Second, by combining this with an algorithm for finding permutations based on maximizing correlations between the activations of matched neurons, we are able to reduce the interpolation barrier for a standard ResNet18 trained on CIFAR-10 to 1.5% absolute test error. We explore the interaction between our method and the choice of normalization layer, and demonstrate its robustness across a variety of architectures and training sets.

****

Multi-View Masked Autoencoders for Visual Control

Younggyo Seo,Junsu Kim,Stephen James,Kimin Lee,Jinwoo Shin,Pieter Abbeel

This paper investigates how to leverage data from multiple cameras to learn representations beneficial for visual control. To this end, we present the Multi-View Masked Autoencoder (MV-MAE), a simple and scalable framework for multi-view representation learning. Our main idea is to mask multiple viewpoints from video frames at random and train a video autoencoder to reconstruct pixels of both masked and unmasked viewpoints. This allows the model to learn representations that capture useful information of the current viewpoint but also the cross-view information from different viewpoints. We evaluate MV-MAE on challenging RLBench visual manipulation tasks by training a reinforcement learning agent on top of frozen representations. Our experiments demonstrate that MV-MAE significantly outperforms other multi-view representation learning approaches. Moreover, we show that the number of cameras can differ between the representation learning phase and the behavior learning phase. By training a single-view control agent on top of multi-view representations from MV-MAE, we achieve 62.3% success rate while the single-view representation learning baseline achieves 42.3%.

****

Boosting Adversarial Training with Masked Adaptive Ensemble

Fudong Lin,Xu Yuan,Nian-Feng Tzeng

Adversarial training (AT) can help improve the robustness of a deep neural network (DNN) against potential adversarial attacks by intentionally injecting adversarial examples into the training data, but this way inevitably incurs standard accuracy degradation to some extent, thereby calling for a trade-off between standard accuracy and robustness. Besides, the prominent AT solutions are vulnerable to sparse attacks, due to “robustness overfitting” upon dense attacks, often adopted by AT to produce a threat model. To tackle such shortcomings, this paper proposes a novel framework, including a detector and a classifier bridged by our newly developed adaptive ensemble. Specifically, a Guided Backpropagation-based detector is designed to sniff adversarial examples, driven by our empirical observation. Meanwhile, a classifier with two encoders is employed for extracting visual representations respectively from clean images and adversarial examples. The adaptive ensemble approach also enables us to mask off a random subset of image patches within input data, eliminating potential adversarial effects when encountering malicious inputs with negligible standard accuracy degradation. As such, our approach enjoys improved robustness, able to withstand both dense and sparse attacks, while maintaining high standard accuracy. Experimental results exhibit that our detector and classifier outperform their state-of-the-art counterparts, in terms of detection accuracy, standard accuracy, and adversarial robustness. For example, on CIFAR-10, our detector achieves the best detection accuracy of 99.6% under dense attacks and of 98.5% under sparse attacks. Our classifier achieves the best standard accuracy of 91.2% and the best robustness against dense attack (or sparse attack) of 57.5% (or 54.8%).

****

Diffusion-GAN: Training GANs with Diffusion

Zhendong Wang,Huangjie Zheng,Pengcheng He,Weizhu Chen,Mingyuan Zhou

Generative adversarial networks (GANs) are challenging to train stably, and a pr omising remedy of injecting instance noise into the discriminator input has not been very effective in practice. In this paper, we propose Diffusion-GAN, a novel GAN framework that leverages a forward diffusion chain to generate Gaussian-mixture distributed instance noise. Diffusion-GAN consists of three components, including an adaptive diffusion process, a diffusion timestep-dependent discriminator, and a generator. Both the observed and generated data are diffused by the adaptive diffusion process via different noise-to-data ratios at each timestep. The timestep-dependent discriminator learns to distinguish the diffused real data from the diffused generated data at each diffusion timestep. The generator learns from the discriminator's feedback by backpropagating through the forward diffusion chain, whose length is adaptively adjusted to balance the noise and data levels. We theoretically show that the discriminator's timestep-dependent strategy gives consistent and helpful guidance to the generator, enabling it to match the true data distribution. We demonstrate the advantages of Diffusion-GAN over strong GAN baselines on various datasets, showing that it can produce more realistic images with higher stability and data efficiency than state-of-the-art GANs.

****

Contextual Subspace Approximation with Neural Householder Transforms
Kerrick Johnstonbaugh,Michael Przystupa,Jacob Keller,Martin Jagersand

Choosing an appropriate action representation is an integral part of solving robotic manipulation problems. Published approaches include latent action models which compress the control space into a low dimensional manifold. These involve training a conditional autoencoder, where the current observation and a low-dimensional action are passed through a neural network decoder to compute high dimensional actuation commands. Such models can have a large number of parameters, and can be difficult to interpret from a user perspective. In this work, we propose that similar performance gains in robotics tasks can be achieved by restructuring the neural network to map observations to a basis for a context-dependent linear actuation subspace. This results in an action interface wherein a user’s actions determine a linear combination of a state-conditioned actuation basis. We introduce the Neural Householder Transform (NHT) as a method for computing this basis. Our results show that reinforcement learning agents trained with NHT in kinematic manipulation and locomotion environments tend to be more robust to hyperparameter choice and achieve higher final success rates compared to agents trained with alternative action representations. NHT agents outperformed agents trained with joint velocity/torque actions, agents trained with an SVD actuation basis, and agents trained with a LASER action interface in the WAMWipe, WAMGrasp, and HalfCheetah environments.

****

Mind the Pool: Convolutional Neural Networks Can Overfit Input Size
Bilal Alsallakh,David Yan,Narine Kokhlikyan,Vivek Miglani,Orion Reblitz-Richardson,Pamela Bhattacharya

We demonstrate how convolutional neural networks can overfit the input size: The accuracy drops significantly when using certain sizes, compared with favorable ones. This issue is inherent to pooling arithmetic, with standard downsampling layers playing a major role in favoring certain input sizes and skewing the weights accordingly. We present a solution to this problem by depriving these layers from the arithmetic cues they use to overfit the input size. Through various examples, we show how our proposed spatially-balanced pooling improves the generalization of the network to arbitrary input sizes and its robustness to translational shifts.

****

Towards Unsupervised Time Series Representation Learning: A Decomposition Perspective
Yan Li,Xinjiang Lu,Jingjing Gu,Haishuai Wang,Dejing Dou

Existing contrastive methods of universal time series representation learning mainly rely on distilling invariant patterns at varying scales and building contrastive loss with the help of negative sampling. However, the invariance assumptions may not hold in real-world time-series data, and the infamous negative sampli ng could bring in new biases for representation learning. In this work, we propose a novel contrastive learning approach toward time series representation learning on top of trend-seasonality decomposition, namely TS-DC. TS-DC differentiates itself from prior methods in three folds: 1) a time series decomposition approach is devised to distill different aspects/components of a complex time series; 2) a novel component-wise contrastive loss is proposed in which negative sampling is not necessary; 3) the informative signals of time series can be captured comprehensively by means of adaptive contrasting. Extensive experiments on different public benchmark datasets validate the superior performance of our proposed representation learning method.

****

Reparameterization through Spatial Gradient Scaling

Alexander Detkov,Mohammad Salameh,Muhammad Fetrat,Jialin Zhang,Robin Luwei,SHANGLING JUI,Di Niu

Reparameterization aims to improve the generalization of deep neural networks by transforming a convolution operation into equivalent multi-branched structures during training. However, there exists a gap in understanding how reparameterization may change and benefit learning processes for neural networks. In this paper, we present a novel spatial gradient scaling method to redistribute learning focus among weights in convolutional neural networks. We prove that spatial gradient scaling achieves the same learning dynamics as a branched reparameterization yet without introducing structural changes into the network. We further propose an analytical approach that dynamically learns scalings for each convolutional layer based on the spatial characteristics of its input feature map gauged by mutual information. Experiments on CIFAR-10, CIFAR-100, and ImageNet show that without searching for reparameterized structures, our proposed scaling method outperforms the state-of-the-art reparameterization methods at a lower computational cost.

****

Boomerang: Local sampling on image manifolds using diffusion models

Lorenzo Luzi,Ali Siahkoohi,Paul M Mayer,Josue Casco-Rodriguez,Richard Baraniuk

Diffusion models can be viewed as mapping points in a high-dimensional latent space onto a low-dimensional learned manifold, typically an image manifold. The intermediate values between the latent space and image manifold can be interpreted as noisy images which are determined by the noise scheduling scheme employed during pre-training. We exploit this interpretation to introduce Boomerang, a local image manifold sampling approach using the dynamics of diffusion models. We call it Boomerang because we first add noise to an input image, moving it closer to the latent space, then bring it back to the image space through diffusion dynamics. We use this method to generate images which are similar, but nonidentical, to the original input images on the image manifold. We are able to set how close the generated image is to the original based on how much noise we add. Additionally, the generated images have a degree of stochasticity, allowing us to locally sample as many times as we want without repetition. We show three applications for which Boomerang can be used. First, we provide a framework for constructing privacy-preserving datasets having controllable degrees of anonymity. Second, we show how to use Boomerang for data augmentation while staying on the image manifold. Third, we introduce a framework for image super-resolution with 8x upsampling. Boomerang does not require any modification to the training of diffusion models and can be used with pretrained models on a single, inexpensive GPU.

****

TOWARD RELIABLE NEURAL SPECIFICATIONS

Chuqin Geng,Nham Le,Xiaojie Xu,Zhaoyue Wang,Arie Gurfinkel,Xujie Si

Having reliable specifications is an unavoidable challenge in achieving verifiable correctness, robustness, and interpretability of AI systems. Existing specifications for neural networks are in the flavor of “data as specification”, that is, the local neighborhood centering around a reference input is considered to be correct (or robust). However, our empirical study shows that such specifications fail to certify any test data points, making it impractical for real-world applications. We propose a new family of specifications called “neural representati on as specification”, which uses the intrinsic information of neural networks — neural activation patterns (NAP) rather than input data to specify the correctness and/or robustness of neural network predictions. We present a simple statistical approach to extracting dominant neural activation patterns. We analyze NAPs from a statistical point of view and find that a single NAP can cover a large number of training and testing data points whereas ad hoc data-as-specification can only cover a single training data point and often zero testing data points. To show the effectiveness of discovered NAPs, we formally verify several important properties, such as a particular type of misclassification never happens for a given NAP, and there is no ambiguity among different NAPs. We show that by using NAP, we can verify the prediction of the entire input space, while still recalling 84% of the data. Thus, we argue that using NAPs is a more reliable and extensible specification for neural network verification.

****

A second order regression model shows edge of stability behavior
Atish Agarwala,Jeffrey Pennington,Fabian Pedregosa

Recent studies of learning algorithms have shown that there is a regime with an initial increase in the largest eigenvalue of the loss Hessian (progressive sharpening), followed by a stabilization of the eigenvalue near the maximum value which allows convergence (edge of stability). We consider a class of predictive models that are quadratic in the parameters, which we call second-order regression models. This is in contrast with the neural tangent kernel regime, where the predictive function is linear in the parameters. For quadratic objectives in two dimensions, we prove that this second order regression model exhibits both progressive sharpening and edge of stability behavior. We then show that in higher dimensions, the model shows this behavior generically without the structure of a neural network, due to a non-linearity induced in the learning dynamics. Finally, we show that edge of stability behavior in neural networks is correlated with the behavior in quadratic regression models.

****

Unsupervised Learning for Combinatorial Optimization Needs Meta Learning
Haoyu Peter Wang,Pan Li

A general framework of unsupervised learning for combinatorial optimization (CO) is to train a neural network whose output gives a problem solution by directly optimizing the CO objective. Albeit with some advantages over traditional solvers, current frameworks optimize an averaged performance over the distribution of historical problem instances, which misaligns with the actual goal of CO that looks for a good solution to every future encountered instance. With this observation, we propose a new objective of unsupervised learning for CO where the goal of learning is to search for good initialization for future problem instances rather than give direct solutions. We propose a meta-learning-based training pipeline for this new objective. Our method achieves good performance. We observe that even the initial solution given by our model before fine-tuning can significantly outperform the baselines under various evaluation settings including evaluation across multiple datasets, and the case with big shifts in the problem scale. The reason we conjecture is that meta-learning-based training lets the model be loosely tied to each local optimum for a training instance while being more adaptive to the changes of optimization landscapes across instances.

****

Latent Topology Induction for Understanding Contextualized Representations
Yao Fu,Mirella Lapata

Recently, there has been considerable interests in understanding pretrained language models. This work studies the hidden geometry of the representation space of language models from a unique topological perspective. We hypothesize that there exist a network of latent anchor states summarizing the topology (neighbors and connectivity) of the representation space. we infer this latent network in a fully unsupervised way using a structured variational autoencoder. We show that such network exists in pretrained representations, but not in baseline random or positional embeddings. We connect the discovered topological structure to their linguistic interpretations. In this latent network, leave nodes can be grounded to word surface forms, anchor states can be grounded to linguistic categories, and connections between nodes and states can be grounded to phrase constructions and syntactic templates. We further show how such network evolves as the embeddings become more contextualized, with observational and statistical evidence demonstrating how contextualization helps words “receive meaning” from their topological neighbors via the anchor states. We demonstrate these insights with extensive experiments and visualizations.

****

DyG2Vec: Representation Learning for Dynamic Graphs With Self-supervision
Mohammad Ali Alomrani,Mahdi Biparva,Yingxue Zhang,Mark Coates

The challenge in learning from dynamic graphs for predictive tasks lies in extracting fine-grained temporal motifs from an ever-evolving graph. Moreover, task labels are often scarce, costly to obtain, and highly imbalanced for large dynamic graphs. Recent advances in self-supervised learning on graphs demonstrate great potential, but focus on static graphs. State-of-the-art (SoTA) models for dynamic graphs are not only incompatible with the self-supervised learning (SSL) paradigm but also fail to forecast interactions beyond the very near future. To address these limitations, we present DyG2Vec, an SSL-compatible, efficient model for representation learning on dynamic graphs. DyG2Vec uses a window-based mechanism to generate task-agnostic node embeddings that can be used to forecast future interactions. DyG2Vec significantly outperforms SoTA baselines on benchmark datasets for downstream tasks while only requiring a fraction of the training/inference time. We adapt two SSL evaluation mechanisms to make them applicable to dynamic graphs and thus show that SSL pre-training helps learn more robust temporal node representations, especially for scenarios with few labels.

****

Unsupervised Meta-learning via Few-shot Pseudo-supervised Contrastive Learning
Huiwon Jang,Hankook Lee,Jinwoo Shin

Unsupervised meta-learning aims to learn generalizable knowledge across a distribution of tasks constructed from unlabeled data. Here, the main challenge is how to construct diverse tasks for meta-learning without label information; recent works have proposed to create, e.g., pseudo-labeling via pretrained representations or creating synthetic samples via generative models. However, such a task construction strategy is fundamentally limited due to heavy reliance on the immutable pseudo-labels during meta-learning and the quality of the representations or the generated samples. To overcome the limitations, we propose a simple yet effective unsupervised meta-learning framework, coined Pseudo-supervised Contrast (PsCo), for few-shot classification. We are inspired by the recent self-supervised learning literature; PsCo utilizes a momentum network and a queue of previous batches to improve pseudo-labeling and construct diverse tasks in a progressive manner. Our extensive experiments demonstrate that PsCo outperforms existing unsupervised meta-learning methods under various in-domain and cross-domain few-shot classification benchmarks. We also validate that PsCo is easily scalable to a large-scale benchmark, while recent prior-art meta-schemes are not.

****

PromptBoosting: Black-Box Text Classification with Ten Forward Passes
Bairu Hou,Joe O'Connor,Jacob Andreas,Shiyu Chang,Yang Zhang

We describe PromptBoosting, a query-efficient procedure for building a text classifier from a neural language model (LM) without access to the LM’s parameters, gradients, or hidden representations. This form of “black-box” classifier training has become increasingly important as the cost of training and inference in large-scale LMs grows. But existing black-box LM classifier learning approaches are themselves computationally inefficient, typically specializing LMs to the target task by searching in a large space of (discrete or continuous) prompts using zeroth-order optimization methods. Instead of directly optimizing in prompt space, PromptBoosting obtains a small pool of prompts via a gradient-free approach and then constructs a large pool of weak learners by pairing these prompts with different elements of the LM’s output distribution. These weak learners are then ensembled using the AdaBoost algorithm. The entire learning process requires only a small number of forward passes and no backward pass. Experiments show that P romptBoosting achieves state-of-the-art performance in multiple black-box few-shot classification tasks, and matches or outperforms full fine-tuning in both few-shot and standard learning paradigms, while training 10x faster than existing black-box methods.

****

Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models

Liam H Fowl,Jonas Geiping,Steven Reich,Yuxin Wen,Wojciech Czaja,Micah Goldblum,Tom Goldstein

Privacy is a central tenet of Federated learning (FL), in which a central server trains models without centralizing user data. However, gradient updates used in FL can leak user information. While the most industrial uses of FL are for text applications (e.g. keystroke prediction), the majority of attacks on user privacy in FL have focused on simple image classifiers and threat models that assume honest execution of the FL protocol from the server. We propose a novel attack that reveals private user text by deploying malicious parameter vectors, and which succeeds even with mini-batches, multiple users, and long sequences. Unlike previous attacks on FL, the attack exploits characteristics of both the Transformer architecture and the token embedding, separately extracting tokens and positional embeddings to retrieve high-fidelity text. We argue that the threat model of malicious server states is highly relevant from a user-centric perspective, and show that in this scenario, text applications using transformer models are much more vulnerable than previously thought.

****

Adaptive Optimization in the ∞-Width Limit

Etai Littwin,Greg Yang

Recent works have developed detailed understanding of large neural networks' behaviors via their infinite-width limits, e.g., the neural tangent kernel (NTK) and the feature learning (μ) limits. These theories were developed for stochastic gradient descent. Yet, in practice, all large NN are trained using Adam or other adaptive gradient optimizers (AGO), which are not covered by such previous works. Here, we close this gap via the Tensor Programs framework. Specifically, for deep MLPs, we derive the NTK and μ parametrizations as well as their infinite-width limits. We find 1) The NTK limit of AGO, in contrast to that of SGD, now depends nonlinearly on the loss derivative but nevertheless still fails to learn features; 2) this is fixed by the μ limit of AGO (as in the case of SGD). To obtain these results, we extend the Tensor Programs language with a new instruction that allows one to express the gradient processing done by AGOs.

****

Pyramidal Denoising Diffusion Probabilistic Models

Dohoon Ryu,Jong Chul Ye

Recently, diffusion model have demonstrated impressive image generation performances, and have been extensively studied in various computer vision tasks. Unfortunately, training and evaluating diffusion models consume a lot of time and computational resources. To address this problem, here we present a novel pyramidal diffusion model that can generate high resolution images starting from much coarser resolution images using a {\em single} score function trained with a positional embedding. This enables a neural network to be much lighter and also enables time-efficient image generation without compromising its performances. Furthermore, we show that the proposed approach can be also efficiently used for multi-scale super-resolution problem using a single score function.

****

Guiding Energy-based Models via Contrastive Latent Variables

Hankook Lee,Jongheon Jeong,Sejun Park,Jinwoo Shin

An energy-based model (EBM) is a popular generative framework that offers both explicit density and architectural flexibility, but training them is difficult since it is often unstable and time-consuming. In recent years, various training techniques have been developed, e.g., better divergence measures or stabilization in MCMC sampling, but there often exists a large gap between EBMs and other generative frameworks like GANs in terms of generation quality. In this paper, we p ropose a novel and effective framework for improving EBMs via contrastive representation learning (CRL). To be specific, we consider representations learned by contrastive methods as the true underlying latent variable. This contrastive latent variable could guide EBMs to understand the data structure better, so it can improve and accelerate EBM training significantly. To enable the joint training of EBM and CRL, we also design a new class of latent-variable EBMs for learning the joint density of data and the contrastive latent variable. Our experimental results demonstrate that our scheme achieves lower FID scores, compared to prior-art EBM methods (e.g., additionally using variational autoencoders or diffusion techniques), even with significantly faster and more memory-efficient training. We also show conditional and compositional generation abilities of our latent-variable EBMs as their additional benefits, even without explicit conditional training. The code is available at https://github.com/hankook/CLEL.

****

Deep Watermarks for Attributing Generative Models

Guangyu Nie,Changhoon Kim,Yezhou Yang,Yi Ren

Generative models have enabled the creation of contents that are indistinguishable from those taken from the Nature. Open-source development of such models raised concerns about the risks in their misuse for malicious purposes. One potential risk mitigation strategy is to attribute generative models via watermarking. Current watermarking methods exhibit significant tradeoff between robust attribution accuracy and generation quality, and also lack principles for designing watermarks to improve this tradeoff. This paper investigates the use of latent semantic dimensions as watermarks, from where we can analyze the effects of design variables, including the choice of watermarking dimensions, watermarking strength, and the capacity of watermarks, on the accuracy-quality tradeoff. Compared with previous SOTA, our method requires minimum computation and is more applicable to large-scale models. We use StyleGAN2 and the latent diffusion model to demonstrate the efficacy of our method.

****

Steerable Equivariant Representation Learning

Sangnie Bhardwaj,Willie McClinton,Tongzhou Wang,Guillaume Lajoie,Chen Sun,Phillip Isola,Dilip Krishnan

Pre-trained deep image representations are useful for post-training tasks such as classification through transfer learning, image retrieval, and object detection. Data augmentations are a crucial aspect of pre-training robust representations in both supervised and self-supervised settings. Data augmentations explicitly or implicitly promote \emph{invariance} in the embedding space to the input image transformations. This invariance reduces generalization to those downstream tasks which rely on sensitivity to these particular data augmentations. In this paper, we propose a method of learning representations that are instead \emph{equivariant} to data augmentations. We achieve this equivariance through the use of \emph{steerable} representations. Our representations can be manipulated directly in embedding space via learned linear maps. We demonstrate that our resulting steerable and equivariant representations lead to better performance on transfer learning and robustness: e.g. we improve linear probe top-1 accuracy by between 1\% to 3\% for transfer; and ImageNet-C accuracy by upto 3.4\%. We further show that the steerability of our representations provides significant speedup (nearly $50\times$) for test-time augmentations; by applying a large number of augmentations for out-of-distribution detection, we significantly improve OOD AUC on the ImageNet-C dataset over an invariant representation.

****

Differentially Private Diffusion Models

Tim Dockhorn,Tianshi Cao,Arash Vahdat,Karsten Kreis

While modern machine learning models rely on increasingly large training datasets, data is often limited in privacy-sensitive domains. Generative models trained with differential privacy (DP) on sensitive data can sidestep this challenge, providing access to synthetic data instead. However, training DP generative models is highly challenging due to the noise injected into training to enforce DP. We propose to leverage diffusion models (DMs), an emerging class of deep generati ve models, and introduce Differentially Private Diffusion Models (DPDMs), which enforce privacy using differentially private stochastic gradient descent (DP-SGD). We motivate why DP-SGD is well suited for training DPDMs, and thoroughly investigate the DM parameterization and the sampling algorithm, which turn out to be crucial ingredients in DPDMs. Furthermore, we propose noise multiplicity, a simple yet powerful modification of the DM training objective tailored to the DP setting to boost performance. We validate our novel DPDMs on widely-used image generation benchmarks and achieve state-of-the-art (SOTA) performance by large margins. For example, on MNIST we improve the SOTA FID from 48.4 to 5.01 and downstream classification accuracy from 83.2% to 98.1% for the privacy setting DP-$(\varepsilon=10, \delta=10^{-5})$. Moreover, on standard benchmarks, classifiers trained on DPDM-generated synthetic data perform on par with task-specific DP-SGD-trained classifiers, which has not been demonstrated before for DP generative models.

****

Outlier-Robust Group Inference via Gradient Space Clustering

Yuchen Zeng,Kristjan Greenewald,Kangwook Lee,Justin Solomon,Mikhail Yurochkin

Traditional machine learning models focus on achieving good performance on the overall training distribution, but they often underperform on minority groups. Existing methods can improve the worst-group performance, but they can have several limitations: (i) they require group annotations, which are often expensive and sometimes infeasible to obtain, and/or (ii) they are sensitive to outliers. Most related works fail to solve these two issues simultaneously as they focus on conflicting perspectives of minority groups and outliers. We address the problem of learning group annotations in the presence of outliers by clustering the data in the space of gradients of the model parameters. We show that data in the gradient space has a simpler structure while preserving information about minority groups and outliers, making it suitable for standard clustering methods like DBSCAN. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art both in terms of group identification and downstream worst-group performance.

****

Broken Neural Scaling Laws

Ethan Caballero,Kshitij Gupta,Irina Rish,David Krueger

We present a smoothly broken power law functional form (referred to by us as a broken neural scaling law (BNSL)) that accurately models and extrapolates the scaling behaviors of deep neural networks (i.e. how the evaluation metric of interest varies as the amount of compute used for training, number of model parameters, training dataset size, or upstream performance varies) for various architectures and for each of various tasks within a large and diverse set of upstream and downstream tasks, in zero-shot, prompted, and fine-tuned settings. This set includes large-scale vision, language, audio, video, diffusion, generative modeling, multimodal learning, contrastive learning, AI alignment, robotics, out-of-distribution generalization, continual learning, uncertainty estimation / calibration, adversarial robustness, molecules, computer programming/coding, math word problems, arithmetic, unsupervised/self-supervised learning, and reinforcement learning (single agent and multi-agent). When compared to other functional forms for neural scaling behavior, this functional form yields extrapolations of scaling behavior that are considerably more accurate on this set. Moreover, this functional form accurately models and extrapolates scaling behavior that other functional forms are incapable of expressing such as the non-monotonic transitions present in the scaling behavior of phenomena such as double descent and the delayed, sharp inflection points present in the scaling behavior of tasks such as arithmetic. Lastly, we use this functional form to glean insights about the limit of the predictability of scaling behavior. See arXiv for longer version of this paper. Code is available at https://github.com/ethancaballero/broken_neural_scaling_laws

****

Learning to perceive objects by prediction

Tushar Arora,JOHN DAY,Li Erran Li,Ming Bo Cai

The representation of objects is the building block of higher-level concepts. Infants develop the notion of objects without supervision, for which the prediction error of future sensory input is likely a major teaching signal. We assume that the goal of representing objects distinctly is to allow the prediction of the coherent motion of all parts of an object independently from the background while keeping track of relatively fewer parameters of the object's motion. To realize this, we propose a framework to extract object-centric representations from single 2D images by learning to predict future scenes containing moving objects. The model learns to explicitly infer objects' locations in a 3D environment, generate 2D segmentation masks of objects, and perceive depth. Importantly, the model requires no supervision or pre-training but assumes rigid-body motion and only needs the observer's self-motion at training time. Further, by evaluating on a new synthetic dataset with more complex textures of objects and background, we found our model overcomes the reliance on clustering colors for segmenting objects, which is a limitation for previous models not using motion information. Our work demonstrates a new approach to learning symbolic representation grounded in sensation and action.
****
Avoiding spurious correlations via logit correction
Sheng Liu,Xu Zhang,Nitesh Sekhar,Yue Wu,Prateek Singhal,Carlos Fernandez-Granda
Empirical studies suggest that machine learning models trained with empirical risk minimization (ERM) often rely on attributes that may be spuriously correlated with the class labels. Such models typically lead to poor performance during inference for data lacking such correlations. In this work, we explicitly consider a situation where potential spurious correlations are present in the majority of training data. In contrast with existing approaches, which use the ERM model outputs to detect the samples without spurious correlations and either heuristically upweight or upsample those samples, we propose the logit correction (LC) loss, a simple yet effective improvement on the softmax cross-entropy loss, to correct the sample logit. We demonstrate that minimizing the LC loss is equivalent to maximizing the group-balanced accuracy, so the proposed LC could mitigate the negative impacts of spurious correlations. Our extensive experimental results further reveal that the proposed LC loss outperforms state-of-the-art solutions on multiple popular benchmarks by a large margin, an average 5.5% absolute improvement, without access to spurious attribute labels. LC is also competitive with oracle methods that make use of the attribute labels.
****
LEARNING CONTEXT-AWARE ADAPTIVE SOLVERS TO ACCELERATE QUADRATIC PROGRAMMING
Haewon Jung,Junyoung Park,Jinkyoo Park
Quadratic programming (QP) is an important sub-field of mathematical optimization. The alternating direction method of multipliers (ADMM) is a successful method to solve QP. Even though ADMM shows promising results in solving various types of QP, its convergence speed is known to be highly dependent on the step-size parameter ρ. Due to the absence of a general rule for setting ρ, it is often tuned manually or heuristically. In this paper, we propose CA-ADMM (Context-aware Adaptive ADMM)) which learns to adaptively adjust ρ to accelerate ADMM. CA-ADMM extracts the spatio-temporal context, which captures the dependency of the primal and dual variables of QP and their temporal evolution during the ADMM iterations. CA-ADMM chooses ρ based on the extracted context. Through extensive numerical experiments, we validated that CA-ADMM effectively generalizes to unseen QP problems with different sizes and classes (i.e., having different QP parameter structures). Furthermore, we verified that CA-ADMM could dynamically adjust ρ considering the stage of the optimization process to accelerate the convergence speed further.
****
Learning Latent Structural Causal Models
Jithendaraa Subramanian,Yashas Annadani,Ivaxi Sheth,Nan Rosemary Ke,Tristan Deleu,Stefan Bauer,Derek Nowrouzezahrai,Samira Ebrahimi Kahou
Causal learning has long concerned itself with the accurate recovery of underlying causal mechanisms. Such causal modelling enables better explanations of out-o

f-distribution data. Prior works on causal learning assume that the high-level causal variables are given. However, in machine learning tasks, one often operates on low-level data like image pixels or high-dimensional vectors. In such settings, the entire Structural Causal Model (SCM) -- structure, parameters, \textit{and} high-level causal variables -- is unobserved and needs to be learnt from low-level data. We treat this problem as Bayesian inference of the latent SCM, given low-level data. For linear Gaussian additive noise SCMs, we present a tractable approximate inference method which performs joint inference over the causal variables, structure and parameters of the latent SCM from random, known interventions. Experiments are performed on synthetic datasets and a causally generated image dataset to demonstrate the efficacy of our approach. We also perform image generation from unseen interventions, thereby verifying out of distribution generalization for the proposed causal model.

****

Pre-Training for Robots: Leveraging Diverse Multitask Data via Offline Reinforcement Learning

Aviral Kumar,Anikait Singh,Frederik Ebert,Yanlai Yang,Chelsea Finn,Sergey Levine

Recent progress in deep learning highlights the tremendous potential of utilizing diverse datasets for achieving effective generalization and makes it enticing to consider leveraging broad datasets for attaining more robust generalization in robotic learning as well. However, in practice we likely will want to learn a new skill in a new environment that is unlikely to be contained in the prior data. Therefore we ask: how can we leverage existing diverse offline datasets in combination with small amounts of task-specific data to solve new tasks, while still enjoying the generalization benefits of training on large amounts of data? In this paper, we demonstrate that end-to-end offline RL can be an effective approach for doing this, without the need for any representation learning or vision-based pre-training. We present pre-training for robots (PTR), a framework based on offline RL that attempts to effectively learn new tasks by combining pre-training on existing robotic datasets with rapid fine-tuning on a new task, with as a few as 10 demonstrations. At its core, PTR applies an existing offline RL method such as conservative Q-learning (CQL), but extends it to include several crucial design decisions that enable PTR to actually work and outperform a variety of prior methods. To the best of our knowledge, PTR is the first offline RL method that succeeds at learning new tasks in a new domain on a real WidowX robot with as few as 10 task demonstrations, by effectively leveraging an existing dataset of diverse multi-task robot data collected in a variety of toy kitchens. We present an accompanying overview video at https://www.youtube.com/watch?v=yAWgyLJD5lY&ab_channel=PTRICLR

****

Safe Exploration Incurs Nearly No Additional Sample Complexity for Reward-Free RL

Ruiquan Huang,Jing Yang,Yingbin Liang

Reward-free reinforcement learning (RF-RL), a recently introduced RL paradigm, relies on random action-taking to explore the unknown environment without any reward feedback information. While the primary goal of the exploration phase in RF-RL is to reduce the uncertainty in the estimated model with minimum number of trajectories, in practice, the agent often needs to abide by certain safety constraint at the same time. It remains unclear how such safe exploration requirement would affect the corresponding sample complexity in order to achieve the desired optimality of the obtained policy in planning. In this work, we make a first attempt to answer this question. In particular, we consider the scenario where a safe baseline policy is known beforehand, and propose a unified Safe reWard-frEe ExploraTion (SWEET) framework. We then particularize the SWEET framework to the tabular and the low-rank MDP settings, and develop algorithms coined Tabular-SWEET and Low-rank-SWEET, respectively. Both algorithms leverage the concavity and continuity of the newly introduced truncated value functions, and are guaranteed to achieve zero constraint violation during exploration with high probability. Furthermore, both algorithms can provably find a near-optimal policy subject to any constraint in the planning phase. Remarkably, the sample complexities under both algorithms match or even outperform the state of the art in their constraint-free counterparts up to some constant factors, proving that safety constraint hardly increases the sample complexity for RF-RL.

****

S^6-DAMON: Bridging Self-Supervised Speech Models and Real-time Speech Recognition

Yonggan Fu,Zhifan Ye,Shunyao Zhang,Jiayi Yuan,Zhongzhi Yu,Yingyan Lin

There has been an growing demand for deep neural network (DNN) powered automatic speech recognition (ASR) on mobile platforms for real-time speech recognition. However, ubiquitous on-device ASR systems are still hindered by two bottlenecks: (1) the lack of large-scale transcribed speech data especially for low-resource spoken languages and (2) the large gap between DNNs' prohibitive complexity and mobiles' limited resources. In parallel, speech models pretrained via self-supervised learning (SSL) have emerged to reduce the reliance on the availability of transcribed speech data, which however further enlarges the efficiency gap because they often adopt large transformers to ensure expressive speech representations. Thus, it is highly desired to trim down the complexity of speech SSL models to enable real-time on-device ASR. This is particularly challenging since only structured sparsity can favor hardware efficiency in commercial devices, under which the speech representation learned by SSL could easily be demolished. To this end, we develop a framework dubbed S^6-DAMON to pursue structured sparsity in speech SSL models via data-model co-compression. On the data side, leveraging both the duration of each phoneme and the pauses between the words/phonemes of human utterances, we propose a salient audio token detector, dubbed SALAD, to remove input audio tokens that are redundant; On the model side, we identify that the failure of the SOTA ASR pruning method under structured sparsity is caused by the sparsity discrepancy between finetuning/deployment and their limited learnability of sparsity distributions, and then tackle it via a new ASR pruning pipeline dubbed SAFARI, which adopts a three-step pipeline - sparsify, finetune, and adjust sparsity. Extensive experiments validate that S^6-DAMON can enable real-time ASR with limited transcribed speech data requirements while maintaining decent recognition performance. All source codes will be released upon acceptance.

****

Teaching Algorithmic Reasoning via In-context Learning

Hattie Zhou,Azade Nova,Aaron Courville,Hugo Larochelle,Behnam Neyshabur,Hanie Sedghi

Large language models (LLMs) have shown increasing in-context learning capabilities through scaling up model and data size. Despite this progress, LLMs are still unable to solve algorithmic reasoning problems. While providing a rationale with the final answer has led to further improvements in multi-step reasoning problems, Anil et al. 2022 showed that even simple algorithmic reasoning tasks such as parity are far from solved. In this work, we identify and study four key stages for successfully teaching algorithmic reasoning to LLMs: (1) formulating algorithms as skills, (2) teaching multiple skills simultaneously (skill accumulation), (3) teaching how to combine skills (skill composition) and (4) teaching how to use skills as tools. We show that it is possible to teach algorithmic reasoning to LLMs via in-context learning, which we refer to as Algorithmic Prompting. We evaluate our approach on a variety of arithmetic and quantitative reasoning tasks, and demonstrate significant boosts in performance over existing prompting techniques. In particular, for long parity, addition, multiplication and subtraction and parity tasks, we achieve an error reduction of approximately 10x, 9x, 5x and 2x respectively compared to the best available baselines.

****

Offline Q-learning on Diverse Multi-Task Data Both Scales And Generalizes

Aviral Kumar,Rishabh Agarwal,Xinyang Geng,George Tucker,Sergey Levine

The potential of offline reinforcement learning (RL) is that high-capacity models trained on large, heterogeneous datasets can lead to agents that generalize broadly, analogously to similar advances in vision and NLP. However, recent works argue that offline RL methods encounter unique challenges to scaling up model capacity. Drawing on the learnings from these works, we re-examine previous design choices and find that with appropriate choices: ResNets, cross-entropy based distributional backups, and feature normalization, offline Q-learning algorithms exhibit strong performance that scales with model capacity. Using multi-task Atari as a testbed for scaling and generalization, we train a single policy on 40 games with near-human performance using up-to 80 million parameter networks, finding that model performance scales favorably with capacity. In contrast to prior work, we extrapolate beyond dataset performance even when trained entirely on a large (400M transitions) but highly suboptimal dataset (51% human-level performance). Compared to return-conditioned supervised approaches, offline Q-learning scales similarly with model capacity and has better performance, especially when the dataset is suboptimal. Finally, we show that offline Q-learning with a diverse dataset is sufficient to learn powerful representations that facilitate rapid transfer to novel games and fast online learning on new variations of a training game, improving over existing state-of-the-art representation learning approaches.

****

Disentangled Conditional Variational Autoencoder for Unsupervised Anomaly Detection

Asif Ahmed Neloy,Maxime Turgeon

Recently, generative models have shown promising performance in anomaly detection tasks. Specifically, autoencoders learn representations of high-dimensional data, and their reconstruction ability can be used to assess whether a new instance is likely to be anomalous. However, the primary challenge of unsupervised anomaly detection (UAD) is in learning appropriate disentangled features and avoiding information loss, while incorporating known sources of variation to improve the reconstruction. In this paper, we propose a novel architecture of generative autoencoder by combining the frameworks of β-VAE, conditional variational autoencoder (CVAE), and the principle of total correlation (TC). We show that our architecture improves the disentanglement of latent features, optimizes TC loss more efficiently, and improves the ability to detect anomalies in an unsupervised manner with respect to high-dimensional instances, such as in imaging datasets. Through both qualitative and quantitative experiments on several benchmark datasets, we demonstrate that our proposed method excels in terms of both anomaly detection and capturing disentangled features. Our analysis underlines the importance of learning disentangled features for UAD tasks.

****

Diffusion-based Image Translation using disentangled style and content representation

Gihyun Kwon,Jong Chul Ye

Diffusion-based image translation guided by semantic texts or a single target image has enabled flexible style transfer which is not limited to the specific domains.

Unfortunately, due to the stochastic nature of diffusion models, it is often difficult to maintain the original content of the image during the reverse diffusion.

To address this, here we present a novel diffusion-based unsupervised image translation method, dubbed as DiffuseIT, using disentangled style and content representation.

Specifically, inspired by the slicing Vision Transformer, we extract intermediate keys of multihead self attention layer from ViT model and used them as the content preservation loss. Then, an image guided style transfer is performed by matching the [CLS] classification token from the denoised samples and target image, whereas additional CLIP loss is used for the text-driven style transfer.

To further accelerate the semantic change during the reverse diffusion, we also propose a novel semantic divergence loss and resampling strategy.

Our experimental results show that the proposed method outperforms state-of-the-art baseline models in both text-guided and image-guided translation tasks.

****

An Analytic Framework for Robust Training of Differentiable Hypothesis

Ramin Barati,Reza Safabakhsh,Mohammad Rahmati

The reliability of a learning model is key to the successful deployment of machine learning in various industries. Creating a robust model, particularly one unaffected by adversarial attacks, requires a comprehensive understanding of the adversarial examples phenomenon. However, it is difficult to describe the phenomenon due to the complicated nature of the problems in machine learning. Consequently, many studies investigate the phenomenon by proposing a simplified model of how adversarial examples occur and validate it by predicting some aspect of the phenomenon. While these studies cover many different characteristics of the adversarial examples, they have not reached a holistic approach to the geometric and analytic modeling of the phenomenon. We observe the phenomenon in many applications of machine learning, and its effects seems to be independent of the choice of the hypothesis class. In this paper, we propose a formalization of robustness in learning theoretic terms and give a geometrical description of the phenomenon in analytic classifiers. We then utilize the proposal to devise a robust classification learning rule for differentiable hypothesis classes and showcase our framework on synthetic and real-world data.

****

Federated Learning with Heterogeneous Label Noise: A Dual Structure Approach

Xinyuan Ji,Zhaowei Zhu,Wei Xi,Olga Gadyatskaya,Zonglin Di,Yang Liu

The performance of federated learning relies heavily on the label quality of each distributed client. In this paper, we consider a federated learning setting with heterogeneous label noise, where each local client might observe training labels with heterogeneous noise rates, which may even drawn from different subsets of the label space. The above high heterogeneity poses challenges for applying the existing label noise learning approaches to each client locally. We formalize the study of federated learning from heterogeneous label noise by firstly identifying two promising label noise generation models. Then, we propose a dual structure approach named FedDual. Intuitively, if there exists a model that filters out the wrongly labeled instances from the local dataset, the effect of label noise can be mitigated. Considering the heterogeneity of local datasets, in addition to the globally shared model, each client in FedDual maintains a local and personalized denoising model. The personalized denoising models can combine information from the global model or other pre-trained models to ensure the performance of denoising. Under this framework, we instantiate our approach with several local sample cleaning methods. We present substantial experiments on MNIST, CIFAR10, and CIFAR100 to demonstrate that FedDual can effectively recognize heterogeneous label noise in different clients and improve the performance of the aggregated model.

****

Mixture of Quantized Experts (MoQE): Complementary Effect of Low-bit Quantization and Robustness

Young Jin Kim,Raffy Fahim,Hany Hassan

Large Mixture of Experts (MoE) models could achieve state-of-the-art quality on various language tasks, including machine translation task, thanks to the efficient model scaling capability with expert parallelism. However, it has brought a fundamental issue of larger memory consumption at deployment time. Furthermore, this results in significant inference speed degradation at auto-regressive decoding steps due to the increased memory transfers. In this paper, we propose a simple weight-only quantization method using ultra low-bit such as 2-bit, 3-bit and 4-bits to effectively mitigate the increased memory and latency issues of MoE models. We show that low-bit quantization together with the MoE architecture delivers a reliable model performance while reducing the memory size significantly even without any additional training. Especially, expert layers in MoE models are much more robust to the quantization than conventional feedforward networks (FFN) layers. In our comprehensive analysis, we show that MoE models with 2-bit and 80\% sparse expert weights can deliver better model performance than the dense model trained on the same dataset. We present how quantization of different parts of models affects the performance with various experiments using a large MoE model (5.3 B). As a result of low-bit quantization, we show the model size can be reduced by 4.9X smaller than the original half precision floating point (fp16)

MoE model. This cuts down the model size of 5.3B parameters from 8.4x of the dense model to only 1.7x of the dense model after 2-bit quantization. It still preserves 1.88\% higher accuracy than the dense model. Combined with an optimized GPU runtime implementation, it also achieves 2.7X speed-up which is even slightly faster than the FLOPs equivalent dense model.

****

Implicit Regularization for Group Sparsity

Jiangyuan Li,Thanh V Nguyen,Chinmay Hegde,Raymond K. W. Wong

We study the implicit regularization of gradient descent towards structured sparsity via a novel neural reparameterization, which we call a diagonally grouped linear neural network. We show the following intriguing property of our reparameterization: gradient descent over the squared regression loss, without any explicit regularization, biases towards solutions with a group sparsity structure. In contrast to many existing works in understanding implicit regularization, we prove that our training trajectory cannot be simulated by mirror descent. We analyze the gradient dynamics of the corresponding regression problem in the general noise setting and obtain minimax-optimal error rates. Compared to existing bounds for implicit sparse regularization using diagonal linear networks, our analysis with the new reparameterization shows improved sample complexity. In the degenerate case of size-one groups, our approach gives rise to a new algorithm for sparse linear regression. Finally, we demonstrate the efficacy of our approach with several numerical experiments.

****

HesScale: Scalable Computation of Hessian Diagonals

Mohamed Elsayed,A. Rupam Mahmood

Second-order optimization uses curvature information about the objective function, which can help in faster convergence. However, such methods typically require expensive computation of the Hessian matrix, preventing their usage in a scalable way. The absence of efficient ways of computation drove the most widely used methods to focus on first-order approximations that do not capture the curvature information. In this paper, we develop \textit{HesScale}, a scalable approach to approximating the diagonal of the Hessian matrix, to incorporate second-order information in a computationally efficient manner. We show that HesScale has the same computational complexity as backpropagation. Our results on supervised classification show that HesScale achieves high approximation accuracy, allowing for scalable and efficient second-order optimization.

****

Divide-and-Cluster: Spatial Decomposition Based Hierarchical Clustering

Narendra Ahuja,Akshat Sharma,Divyam Goel

This paper is about increasing the computational efficiency of clustering algorithms. Many clustering algorithms are based on properties of relative locations of points, globally or locally, e.g., interpoint distances and nearest neighbor distances. This amounts to using a lower dimensional space than the full dimensionality D of the space in which the points are embedded. We present a clustering algorithm, Divide-and-Cluster (DAC), which detects local clusters in small neighborhoods obtained by recursive tessellation of space, and then merges them hierarchically, following the Divide-and-Conquer paradigm. This significantly reduces computation time which may otherwise grow nonlinearly number n of points. We define locality as hypercubical neighborhoods in a recursive hypercubical decomposition of space, represented by a tree. Clusters are detected within each hypercube, and merged with those from neighboring hypercubes while traversing up the tree. We expect DAC to perform better than many other algorithms because (a) as clusters merge into larger clusters (components), their number steadily decreases vs the number of points, and (b) we cluster only neighboring components. The ordering of component appearances also simultaneously yields a cluster hierarchy (tree). Further, our use of small neighborhoods allows piecewise uniform approximation of large, nonuniform, arbitrary shaped clusters, thus avoiding the need for global cluster models. We experimentally verify the correctness of detected clusters on a variety of datasets, posing a variety of challenges, as well as show that DAC’s runtime is significantly better than representative algorithms o f other types, particularly for increasing values of n.

****

Implicit regularization in Heavy-ball momentum accelerated stochastic gradient descent

Avrajit Ghosh,He Lyu,Xitong Zhang,Rongrong Wang

It is well known that the finite step-size (h) in Gradient descent (GD) implicitly regularizes solutions to flatter minimas. A natural question to ask is \textit{Does the momentum parameter β (say) play a role in implicit regularization in Heavy-ball (H.B) momentum accelerated gradient descent (GD+M)?}. To answer this question, first, we show that the trajectory traced by discrete H.B momentum update (GD+M) is $O(h^2)$ close to a continuous trajectory induced by a modified loss, which consists of an original loss and an implicit regularizer. This implicit regularizer for (GD+M) is indeed stronger than that of (GD) by factor of $(\frac{1+\beta}{1-\beta})$, thus explaining why (GD+M) shows better generalization performance and higher test accuracy than (GD). Furthermore, we extend our analysis to stochastic version of gradient descent with momentum (SGD+M) and propose a deterministic continuous trajectory that is $O(h^2)$ close to the discrete update of (SGD+M) in a strong approximation sense. We explore the implicit regularization in (SGD+M) and (GD+M) through a series of experiments validating our theory.

****

ORCA: Interpreting Prompted Language Models via Locating Supporting Evidence in the Ocean of Pretraining Data

Xiaochuang Han,Yulia Tsvetkov

Prompting large pretrained language models leads to strong performance in a variety of downstream tasks. However, it is still unclear from where the model learns task-specific knowledge, especially in zero-shot setups. In this work, we propose a novel method ORCA to identify evidence of the model's task-specific competence in prompt-based learning. Through an instance attribution approach to model interpretability, by iteratively using gradient information related to the downstream task, ORCA locates a very small subset of pretraining data that directly supports the model's predictions in a given task; we call this subset supporting data evidence. We show that supporting data evidence offers new insights about the prompted language models. For example, in the tasks of sentiment analysis and textual entailment, BERT shows a substantial reliance on BookCorpus---the smaller corpus of BERT's two pretraining corpora---as well as on pretraining examples that mask out synonyms to the task labels used in prompts.

****

Real-time variational method for learning neural trajectory and its dynamics

Matthew Dowling,Yuan Zhao,Il Memming Park

Latent variable models have become instrumental in computational neuroscience for reasoning about neural computation. This has fostered the development of powerful offline algorithms for extracting latent neural trajectories from neural recordings. However, despite the potential of real-time alternatives to give immediate feedback to experimentalists, and enhance experimental design, they have received markedly less attention. In this work, we introduce the exponential family variational Kalman filter (eVKF), an online recursive Bayesian method aimed at inferring latent trajectories while simultaneously learning the dynamical system generating them. eVKF works for arbitrary likelihoods and utilizes the constant base measure exponential family to model the latent state stochasticity. We derive a closed-form variational analog to the predict step of the Kalman filter which leads to a provably tighter bound on the ELBO compared to another online variational method. We validate our method on synthetic and real-world data, and, notably, show that it achieves competitive performance.

****

Large Language Models are Human-Level Prompt Engineers

Yongchao Zhou,Andrei Ioan Muresanu,Ziwen Han,Keiran Paster,Silviu Pitis,Harris Chan,Jimmy Ba

By conditioning on natural language instructions, large language models (LLMs) h ave displayed impressive capabilities as general-purpose computers. However, task performance depends significantly on the quality of the prompt used to steer the model, and most effective prompts have been handcrafted by humans. Inspired by classical program synthesis and the human approach to prompt engineering, we propose Automatic Prompt Engineer (APE) for automatic instruction generation and selection. In our method, we treat the instruction as the "program," optimized by searching over a pool of instruction candidates proposed by an LLM in order to maximize a chosen score function. To evaluate the quality of the selected instruction, we evaluate the zero-shot performance of another LLM following the selected instruction. Experiments on 24 NLP tasks show that our automatically generated instructions outperform the prior LLM baseline by a large margin and achieve better or comparable performance to the instructions generated by human annotators on 21/24 tasks. We conduct extensive qualitative and quantitative analyses to explore the performance of APE. We show that APE-engineered prompts can be applied to steer models toward truthfulness and/or informativeness, as well as to improve few-shot learning performance by simply prepending them to standard in-context learning prompts.

****

Do Not Blindly Imitate the Teacher: Loss Perturbation for Knowledge Distillation

Rongzhi Zhang,Jiaming Shen,Tianqi Liu,Jialu Liu,Michael Bendersky,Marc Najork,Chao Zhang

Knowledge distillation (KD) is a popular model compression technique to transfer knowledge from large teacher models to a small student model. Typically, the student learns to imitate the teacher by minimizing the KL divergence of its output distribution with the teacher's output distribution. We argue that such a learning objective is sub-optimal because there exists a discrepancy between the teacher's output distribution and the ground truth label distribution, and forcing the student to blindly imitate the unreliable teacher output distribution leads to inferior performance. To this end, we propose a novel knowledge distillation objective PTLoss by first representing the vanilla KL-based distillation loss function via a Maclaurin series and then perturbing the leading-order terms in this series. This perturbed loss improves the student generalizability by effectively distilling knowledge from a shifted distribution closer to the ground truth data. We also propose a method to compute this shifted teacher distribution, named Proxy Teacher, which enables us to select the perturbation coefficients in PTLoss. We theoretically show the perturbed loss reduces the deviation from the true population risk compared to the vanilla KL-based distillation loss functions. Experiments on three tasks with teachers of different scales show that our method significantly outperforms vanilla distillation loss functions and other perturbation methods.

****

Fast Yet Effective Graph Unlearning through Influence Analysis

Kun Wu,Jie Shen,Yue Ning,Wendy Hui Wang

Recent evolving data privacy policies and regulations have led to increasing interest in the machine unlearning problem. In this paper, we consider Graph Neural Networks (GNNs) as the target model, and study the problem of edge unlearning in GNNs, i.e., learning a new GNN model as if a specified set of edges never existed in the original training graph. Despite its practical importance, the problem remains elusive due to the non-convexity nature of GNNs. Our main technical contribution is three-fold: 1) we cast the problem of edge unlearning as estimating the influence functions of the edges to be removed; 2) we design a computationally and memory efficient algorithm named EraEdge for edge influence estimation and unlearning; 3) under standard regularity conditions, we prove that the sequence of iterates produced by our algorithm converges to the desired model. A comprehensive set of experiments on three prominent GNN models and four benchmark graph datasets demonstrate that our algorithm achieves significant speed-up gains over retraining from scratch without sacrificing the model accuracy too much. Furthermore, our algorithm outperforms the existing GNN unlearning approach in terms of both training time and accuracy of the target GNN model.

****

Faster Hyperparameter Search for GNNs via Calibrated Dataset Condensation
Mucong Ding,Xiaoyu Liu,Tahseen Rabbani,Teresa Ranadive,Tai-Ching Tuan,Furong Huang

Dataset condensation aims to reduce the computational cost of training multiple models on a large dataset by condensing the training dataset into a small synthetic set. State-of-the-art approaches rely on matching the model gradients for the real and synthetic data and have recently been applied to condense large-scale graphs for node classification tasks. Although dataset condensation may be efficient when training multiple models for hyperparameter optimization, there is no theoretical guarantee on the generalizability of the condensed data: data condensation often generalizes poorly across hyperparameters/architectures in practice, while we find and prove this overfitting is much more severe on graphs. In this paper, we consider a different condensation objective specifically geared towards hyperparameter search. We aim to generate the synthetic dataset so that the validation-performance rankings of the models, with different hyperparameters, on the condensed and original datasets are comparable. We propose a novel hyperparameter-calibrated dataset condensation algorithm, which obtains the synthetic validation data by matching the hyperparameter gradients computed via implicit differentiation and efficient inverse Hessian approximation. HCDC employs a supernet with differentiable hyperparameters, making it suitable for modeling GNNs with widely different convolution filters. Experiments demonstrate that the proposed framework effectively maintains the validation-performance rankings of GNNs and speeds up hyperparameter/architecture search on graphs.

****

Spatiotemporal Modeling of Multivariate Signals with Graph Neural Networks and Structured State Space Models
Siyi Tang,Jared Dunnmon,Liangqiong Qu,Khaled Kamal Saab,Christopher Lee-Messer,Daniel Rubin

Multivariate signals are prevalent in various domains, such as healthcare, transportation systems, and space sciences. Modeling spatiotemporal dependencies in multivariate signals is challenging due to (1) long-range temporal dependencies and (2) complex spatial correlations between sensors. To address these challenges, we propose representing multivariate signals as graphs and introduce GraphS4mer, a general graph neural network (GNN) architecture that captures both spatial and temporal dependencies in multivariate signals. Specifically, (1) we leverage Structured State Spaces model (S4), a state-of-the-art sequence model, to capture long-term temporal dependencies and (2) we propose a graph structure learning layer in GraphS4mer to automatically learn the underlying graph structures in the data. We evaluate our proposed model on three distinct tasks and show that GraphS4mer consistently improves over existing models, including (1) seizure detection from electroencephalography signals, outperforming a previous GNN with self-supervised pretraining by 3.1 points in AUROC; (2) sleep staging from polysomnography signals, a 4.1 points improvement in macro-F1 score compared to existing sleep staging models; and (3) traffic forecasting, reducing MAE by 8.8% compared to existing GNNs and by 1.4% compared to transformer-based models.

****

Pruning Deep Neural Networks from a Sparsity Perspective
Enmao Diao,Ganghua Wang,Jiawei Zhang,Yuhong Yang,Jie Ding,Vahid Tarokh

In recent years, deep network pruning has attracted significant attention in order to enable the rapid deployment of AI into small devices with computation and memory constraints. Pruning is often achieved by dropping redundant weights, neurons, or layers of a deep network while attempting to retain a comparable test performance. Many deep pruning algorithms have been proposed with impressive empirical success. However, existing approaches lack a quantifiable measure to estimate the compressibility of a sub-network during each pruning iteration and thus may under-prune or over-prune the model. In this work, we propose PQ Index (PQI) to measure the potential compressibility of deep neural networks and use this to develop a Sparsity-informed Adaptive Pruning (SAP) algorithm. Our extensive experiments corroborate the hypothesis that for a generic pruning procedure, PQI decreases first when a large model is being effectively regularized and then incr eases when its compressibility reaches a limit that appears to correspond to the beginning of underfitting. Subsequently, PQI decreases again when the model collapse and significant deterioration in the performance of the model start to occur. Additionally, our experiments demonstrate that the proposed adaptive pruning algorithm with proper choice of hyper-parameters is superior to the iterative pruning algorithms such as the lottery ticket-based pruning methods, in terms of both compression efficiency and robustness.

****

High-dimensional Continuum Armed and High-dimensional Contextual Bandit: with Applications to Assortment and Pricing

Junhui Cai,Ran Chen,Martin J. Wainwright,Linda H. Zhao

The bandit problem with high-dimensional continuum arms and high-dimensional contextual covariates is often faced by decision-makers but remains unsolved. Existing bandit algorithms are impracticable due to the complexity of the double-layer high dimensionality. We formulate this problem as a high-dimensional continuum armed contextual bandit with high-dimensional covariates and propose a novel model that captures the effect of the arm and contextual on the reward with a low-rank representation matrix. The representation matrix is endowed with interpretability and predictive power. We further propose an efficient bandit algorithm based on a low-rank matrix estimator with theoretical justifications. The generality of our model allows wide applications including business and healthcare. In particular, we apply our method to assortment and pricing, both of which are important decisions for firms such as online retailers. Our method can solve the assortment-pricing problem simultaneously while most existing methods address them separately. We demonstrate the effectiveness of our method to jointly optimize assortment and pricing for revenue maximization for a giant online retailer.

****

■■What learning algorithm is in-context learning? Investigations with linear models

Ekin Akyürek,Dale Schuurmans,Jacob Andreas,Tengyu Ma,Denny Zhou

Neural sequence models, especially transformers, exhibit a remarkable capacity for in-context learning. They can construct new predictors from sequences of labeled examples $(x, f(x))$ presented in the input without further parameter updates. We investigate the hypothesis that transformer-based in-context learners implement standard learning algorithms implicitly, by encoding context-specific parametric models in their hidden representations, and updating these implicit models as new examples appear in the context. Using linear regression as a model problem, we offer three sources of evidence for this hypothesis. First, we prove by construction that transformers can implement learning algorithms for linear models based on gradient descent and closed-form computation of regression parameters. Second, we show that trained in-context learners closely match the predictors computed by gradient descent, ridge regression, and exact least-squares regression, transitioning between different predictors as transformer depth and dataset noise vary. Third, we present preliminary evidence that in-context learners share algorithmic features with these predictors: learners' late layers encode weight vectors and moment matrices. These results suggest that in-context learning is understandable in algorithmic terms, and that (at least in the linear case) learners may work by rediscovering standard estimation algorithms.

****

Learning to represent and predict evolving visual signals via polar straightening

Pierre-Étienne H Fiquet,Eero P Simoncelli

Observer motion and continuous deformations of objects and textures imbue natural videos with distinct temporal structures, enabling the prediction of future frames from past ones. Conventional methods proceed by estimating local motion, or optic flow, and then using this to predict future frames by warping and copying content. Here, we explore a more direct methodology, in which frames are transformed into an alternative representation where temporal structure and evolution are more readily accessible. As a base case, a rigidly translating pattern can be described in the frequency domain as a linear combination of sinusoids, each with constant amplitude and phase that cycles at a rate proportional to its frequency. This fundamental property of Fourier representation reduces prediction to angular extrapolation. Motivated by the geometry of this well-known case, we formulate a self-supervised learning problem which seeks a transformation of video frames to facilitate next-frame prediction in these natural polar coordinates. We construct a network architecture in which pairs of convolutional channels are used to factorize signals into slowly evolving amplitudes and linearly advancing phases. We train this network to predict future frames, and compare its performance with that of conventional methods using optic flow, and other learned predictive neural networks, evaluated on natural videos from the DAVIS dataset. We find that the polar predictor achieves high prediction performance while remaining interpretable and fast, thereby demonstrating the potential of a flow-free video processing methodology that is trained end-to-end to predict natural video content.

****

On Representation Learning Under Class Imbalance
Ravid Shwartz-Ziv,Micah Goldblum,Yucen Lily Li,C. Bayan Bruss,Andrew Gordon Wilson

Unlike carefully curated academic benchmarks, real-world datasets are often highly class-imbalanced, involving training and test sets which contain few examples from certain minority classes. While there is a common understanding that neural network generalization is negatively impacted by imbalance, the source of this problem and its resolution are unclear. Through extensive empirical investigation, we study foundational learning behaviors for various models such as neural networks, gradient-boosted decision trees, and SVMs across a range of domains and find that (1) contrary to conventional wisdom, re-balancing the training set to include a higher proportion of minority samples degrades performance on imbalanced test sets; (2) minority samples are hard to fit, yet algorithms which fit them, such as oversampling, do not improve generalization. Motivated by the observation that re-balancing class-imbalanced training data is ineffective, we show that several existing techniques for improving representation learning are effective in this setting: (3) self-supervised pre-training is insensitive to imbalance and can be used for feature learning before fine-tuning on labels; (4) Bayesian inference is effective because neural networks are especially underspecified under class imbalance; (5) flatness-seeking regularization pulls decision boundaries away from minority samples, especially when we seek minima that are particularly flat on the minority samples’ loss.

****

Gradient Descent Converges Linearly for Logistic Regression on Separable Data
Kyriakos Axiotis,Maxim Sviridenko

We show that running gradient descent on the logistic regression objective guarantees loss $f(x) \leq 1.1 \cdot f(x^*) + \epsilon$, where the error ϵ decays exponentially with the number of iterations. This is in contrast to the common intuition that the absence of strong convexity precludes linear convergence of first-order methods, and highlights the importance of variable learning rates for gradient descent. For separable data, our analysis proves that the error between the predictor returned by gradient descent and the hard SVM predictor decays as $\mathrm{poly}(1/t)$, exponentially faster than the previously known bound of $O(\log\log t / \log t)$. Our key observation is a property of the logistic loss that we call multiplicative smoothness and is (surprisingly) little-explored: As the loss decreases, the objective becomes (locally) smoother and therefore the learning rate can increase. Our results also extend to sparse logistic regression, where they lead to an exponential improvement of the sparsity-error tradeoff.

****

Interpretable (meta)factorization of clinical questionnaires to identify general dimensions of psychopathology
Ka Chun Lam,Bridget W Mahony,Armin Raznahan,Francisco Pereira

Psychiatry research aims at understanding manifestations of psychopathology in b ehavior, in terms of a small number of latent constructs. These are usually inferred from questionnaire data using factor analysis. The resulting factors and relationship to the original questions are not necessarily interpretable. Furthermore, this approach does not provide a way to separate the effect of confounds from those of constructs, and requires explicit imputation for missing data. Finally, there is no clear way to integrate multiple sets of constructs estimated from different questionnaires. An important question is whether there is a universal, compact set of constructs that would span all the psychopathology issues listed across those questionnaires. We propose a new matrix factorization method designed for questionnaires aimed at promoting interpretability, through bound and sparsity constraints. We provide an optimization procedure with theoretical convergence guarantees, and validate automated methods to detect latent dimensionality on synthetic data. We first demonstrate the method on a commonly used general-purpose questionnaire. We then show it can be used to extract a broad set of 15 psychopathology factors spanning 21 questionnaires from the Healthy Brain Network study. We show that our method preserves diagnostic information against competing methods, even as it imposes more constraints. Finally, we demonstrate that it can be used for defining a short, general questionnaire that allows recovery of those 15 meta-factors, using data more efficiently than other methods.

****

Enhancing Meta Learning via Multi-Objective Soft Improvement Functions

Runsheng Yu,Weiyu Chen,Xinrun Wang,James Kwok

Meta-learning tries to leverage information from similar learning tasks. In the commonly-used bilevel optimization formulation, the shared parameter is learned in the outer loop by minimizing the average loss over all tasks. However, the converged solution may be comprised in that it only focuses on optimizing on a small subset of tasks. To alleviate this problem, we consider meta-learning as a multi-objective optimization (MOO) problem, in which each task is an objective. However, existing MOO solvers need to access all the objectives’ gradients in each iteration, and cannot scale to the huge number of tasks in typical meta-learning settings. To alleviate this problem, we propose a scalable gradient-based solver with the use of mini-batch. We provide theoretical guarantees on the Pareto optimality or Pareto stationarity of the converged solution. Empirical studies on various machine learning settings demonstrate that the proposed method is efficient, and achieves better performance than the baselines, particularly on improving the performance of the poorly-performing tasks and thus alleviating the compromising phenomenon.

****

Discrete Predictor-Corrector Diffusion Models for Image Synthesis

Jose Lezama,Tim Salimans,Lu Jiang,Huiwen Chang,Jonathan Ho,Irfan Essa

We introduce Discrete Predictor-Corrector diffusion models (DPC), extending predictor-corrector samplers in Gaussian diffusion models to the discrete case. Predictor-corrector samplers are a class of samplers for diffusion models, which improve on ancestral samplers by correcting the sampling distribution of intermediate diffusion states using MCMC methods. In DPC, the Langevin corrector, which does not have a direct counterpart in discrete space, is replaced with a discrete MCMC transition defined by a learned corrector kernel. The corrector kernel is trained to make the correction steps achieve asymptotic convergence, in distribution, to the correct marginal of the intermediate diffusion states. Equipped with DPC, we revisit recent transformer-based non-autoregressive generative models through the lens of discrete diffusion, and find that DPC can alleviate the compounding decoding error due to the parallel sampling of visual tokens. Our experiments show that DPC improves upon existing discrete latent space models for class-conditional image generation on ImageNet, and outperforms continuous diffusion models and GANs, according to standard metrics and user preference studies.

****

Instruction-Following Agents with Jointly Pre-Trained Vision-Language Models

Hao Liu,Lisa Lee,Kimin Lee,Pieter Abbeel

Humans are excellent at understanding language and vision to accomplish a wide range of tasks. In contrast, creating general instruction-following embodied agen ts remains a difficult challenge. Prior work that uses pure language-only models lack visual grounding, making it difficult to connect language instructions with visual observations. On the other hand, methods that use pre-trained vision-language models typically come with divided language and visual representations, requiring designing specialized network architecture to fuse them together. We propose a simple yet effective model for robots to solve instruction-following tasks in vision-based environments. Our InstructRL method consists of a multimodal transformer that encodes visual observations and language instructions, and a policy transformer that predicts actions based on encoded representations. The multimodal transformer is pre-trained on millions of image-text pairs and natural language text, thereby producing generic cross-modal representations of observations and instructions. The policy transformer keeps track of the full history of observations and actions, and predicts actions autoregressively. We show that this unified transformer model outperforms all state-of-the-art pre-trained or trained-from-scratch methods in both single-task and multi-task settings. Our model also shows better model scalability and generalization ability than prior work.

****

Infusing Lattice Symmetry Priors in Neural Networks Using Soft Attention Masks

Mattia Atzeni,Mrinmaya Sachan,Andreas Loukas

Infusing inductive biases and knowledge priors in artificial neural networks is a promising approach for achieving sample efficiency in current deep learning models. Core knowledge priors of human intelligence have been studied extensively in developmental science and recent work has postulated the idea that research on artificial intelligence should revolve around the same basic priors. As a step towards this direction, in this paper, we introduce LatFormer, a model that incorporates lattice geometry and topology priors in attention masks.

Our study of the properties of these masks motivates a modification to the standard attention mechanism, where attention weights are scaled using soft attention masks generated by a convolutional neural network. Our experiments on ARC and on synthetic visual reasoning tasks show that LatFormer requires 2-orders of magnitude fewer data than standard attention and transformers in these tasks. Moreover, our results on ARC tasks that incorporate geometric priors provide preliminary evidence that deep learning can tackle this complex dataset, which is widely viewed as an important open challenge for AI research.

****

Counterfactual Vision-Language Data Synthesis with Intra-Sample Contrast Learning

Zhecan Wang,Yicheng He,Wenhao Li,Haoxuan You,Long Chen,Noel C Codella,Yulei Niu,Kai-Wei Chang,Shih-Fu Chang

Existing Visual Learning (VL) benchmarks often contain exploitative biases. Most former works only attempted to mitigate biases in semantically low-level and conventional visual-question-answering typed datasets like VQA and GQA. However, these methods cannot generalize to recently emerging highly semantic VL datasets like VCR and are also difficult to scale due to many severe problems like high-cost labors, drastically disrupting the data distribution\textit{, etc.}To resolve those problems and also address other unique biases on VCR-like datasets, we first conduct in-depth analysis and identify important biases in VCR dataset. We further propose a generalized solution that synthesizes counterfactual image and text data based on the original query's semantic focus while producing less distortion to the data distribution. To utilize our synthesized data, we also design an innovative intra-sample contrastive training strategy to assist QA learning in Visual Commonsense Reasoning (VCR). Moreover, our synthesized VL data also serve as a highly-semantic debiased benchmark for evaluating future VL models' robustness. Extensive experiments show that our proposed synthesized data and training strategy improve existing VL models' performances on both the original VCR dataset and our proposed debiased benchmark.

****

OPTQ: Accurate Quantization for Generative Pre-trained Transformers

Elias Frantar,Saleh Ashkboos,Torsten Hoefler,Dan Alistarh

Generative Pre-trained Transformer models, known as GPT or OPT, set themselves a part through breakthrough performance across complex language modelling tasks, but also by their extremely high computational and storage costs. Specifically, due to their massive size, even inference for large, highly-accurate GPT models may require multiple performant GPUs, which limits the usability of such models. While there is emerging work on relieving this pressure via model compression, the applicability and performance of existing compression techniques is limited by the scale and complexity of GPT models. In this paper, we address this challenge, and propose OPTQ, a new one-shot weight quantization method based on approximate second-order information, that is both highly-accurate and highly-efficient. Specifically, OPTQ can quantize GPT models with 175 billion parameters in approximately four GPU hours, reducing the bitwidth down to 3 or 4 bits per weight, with negligible accuracy degradation relative to the uncompressed baseline. Our method more than doubles the compression gains relative to previously-proposed one-shot quantization methods, preserving accuracy, allowing us for the first time to execute an 175 billion-parameter model inside a single GPU for generative inference. Moreover, we also show that our method can still provide reasonable accuracy in the extreme quantization regime, in which weights are quantized to 2-bit or even ternary quantization levels.

We show experimentally that these improvements can be leveraged for end-to-end inference speedups over FP16, of around 3.25x when using high-end GPUs (NVIDIA A100) and 4.5x when using more cost-effective ones (NVIDIA A6000). The implementation is available at https://github.com/IST-DASLab/gptq.

****

ConserWeightive Behavioral Cloning for Reliable Offline Reinforcement Learning

Tung Nguyen,Qinqing Zheng,Aditya Grover

The goal of offline reinforcement learning (RL) is to learn near-optimal policies from static logged datasets, thus sidestepping expensive online interactions. Behavioral cloning (BC) provides a straightforward solution to offline RL by mimicking offline trajectories via supervised learning. Recent advances~\cite{chen2021decision, janner2021offline, emmons2021rvs} have shown that by conditioning on desired future returns, BC can perform competitively to their value-based counterparts, while enjoying much more simplicity and training stability. However, the distribution of returns in the offline dataset can be arbitrarily skewed and suboptimal, which poses a unique challenge for conditioning BC on expert returns at test-time. We propose ConserWeightive Behavioral Cloning (\name), a simple and effective method for improving the performance of conditional BC for offline RL with two key components: trajectory weighting and conservative regularization. Trajectory weighting addresses the bias-variance tradeoff in conditional BC and provides a principled mechanism to learn from both low return trajectories (typically plentiful) and high return trajectories (typically few). Further, we analyze the notion of conservatism in existing BC methods, and propose a novel conservative regularizer that explicitly encourages the policy to stay close to the data distribution. The regularizer helps achieve more reliable performance, and removes the need for ad-hoc tuning of the conditioning value during evaluation. We instantiate \name{} in the context of Reinforcement Learning via Supervised Learning (RvS)~\cite{emmons2021rvs} and Decision Transformer (DT)~\citep{chen2021decision}, and empirically show that it significantly boosts the performance and stability of prior methods on various offline RL benchmarks.

****

ADVL: Adaptive Distillation for Vision-Language Tasks

Zhecan Wang,Noel C Codella,Haoxuan You,Long Chen,Yen-Chun Chen,Yulei Niu,Jianwei Yang,Luowei Zhou,Lu Yuan,Kai-Wei Chang,Shih-Fu Chang

Large-scale image-text pairs, such as image-captions and image-phrases, enable the strong representation of vision-language (VL) models. Nevertheless, they lose diversity and complexity due to the constraints in collecting data. Meanwhile, models pre-trained with image-only or text-only data (we call them unimodal pretrained models) continue to flourish and impress the community. Compared to image-text pairs, unimodal data has less constraints during the collection process resulting in more diverse styles. A natural question is how to leverage unimodal pretrained models to benefit downstream VL tasks? Most existing works focus on fu sing VL information in the expensive pre-training stage. They directly plug in unimodal pre-trained encoders into a VL framework and redo an additional pre-training step on paired image-text data. This causes additional computation expense and the unimodal pretrained knowledge might be forgotten. In this paper, we take a different route and investigate how to fuse VL information in the finetuning stage oaly. To directly transfer pretrained knowledge from unimodal models to help downstream VL tasks, we propose ADVL, which avoids redoing any pre-training step and is generalizable to be applied of top of various VL models. To comprehensively demonstrate the effectiveness of ADVL, we conduct evaluation across three mostly recognized highly semantic VL benchmarks: VCR, VQA, and SNLI-VE under three settings, low-shot, full-shot and domainshifted settings. Results show that ADVL consistently improves the performance with different VL base models across all settings. It even achieves state-of-theart (SOTA) performance on VCR among models pre-trained with image-text data and delivers competitive results on VQA and SNLI-VE, Based on our analysis, we also discover that ADVL can improve the robustness of VL models and regulate them to better use vision information.

****

A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution

Sungyoon Lee,Cheongjae Jang

For full-batch gradient descent (GD), it has been empirically shown that the sharpness, the top eigenvalue of the Hessian, increases and then hovers above $2/\text{(learning rate)}$, and this is called ``the edge of stability'' phenomenon. However, it is unclear why the sharpness is somewhat larger than $2/\text{(learning rate)}$ and how this can be extended to general mini-batch stochastic gradient descent (SGD). We propose a new sharpness measure (interaction-aware-sharpness) aware of the \emph{interaction} between the batch gradient distribution and the loss landscape geometry. This leads to a more refined and general characterization of the edge of stability for SGD. Moreover, based on the analysis of a concentration measure of the batch gradient, we propose a more accurate scaling rule, Linear and Saturation Scaling Rule (LSSR), between batch size and learning rate.

****

$\mathrm{SE}(3)$-Equivariant Attention Networks for Shape Reconstruction in Function Space

Evangelos Chatzipantazis,Stefanos Pertigkiozoglou,Edgar Dobriban,Kostas Daniilidis

We propose a method for 3D shape reconstruction from unoriented point clouds. Our method consists of a novel SE(3)-equivariant coordinate-based network (TF-ONet), that parametrizes the occupancy field of the shape and respects the inherent symmetries of the problem. In contrast to previous shape reconstruction methods that align the input to a regular grid, we operate directly on the irregular point cloud. Our architecture leverages equivariant attention layers that operate on local tokens. This mechanism enables local shape modelling, a crucial property for scalability to large scenes. Given an unoriented, sparse, noisy point cloud as input, we produce equivariant features for each point. These serve as keys and values for the subsequent equivariant cross-attention blocks that parametrize the occupancy field. By querying an arbitrary point in space, we predict its occupancy score. We show that our method outperforms previous SO(3)-equivariant methods, as well as non-equivariant methods trained on SO(3)-augmented datasets. More importantly, local modelling together with SE(3)-equivariance create an ideal setting for SE(3) scene reconstruction. We show that by training only on single, aligned objects and without any pre-segmentation, we can reconstruct novel scenes containing arbitrarily many objects in random poses without any performance loss.

****

PBES: PCA Based Exemplar Sampling Algorithm for Continual Learning

Sahil Nokhwal,Nirman Kumar

Traditional machine learning is both data and computation intensive. The most powerful models require huge quantities of data to train and the training is highly time-consuming. In the streaming or incremental model of machine learning, the data is received and processed in a streaming manner, i.e., the entire data stream is not stored, and the models are updated incrementally. While this is closer to the learning process of humans, a common problem associated with this is “catastrophic forgetting” (CF), i.e., because the entire data is not stored, but just a sketch of it, as more and more data arrives, the older data has invariably a smaller representation in the stored sketch, and this causes models to perform badly on tasks that are closer to older data. One of the approaches to solve this problem stores an “exemplar set” of data items from the stream – but this raises the central question: how to choose which items to store? Current approaches to solve this are based on herding, which is a way to select a random looking sample by a deterministic algorithm. We propose a novel selection approach based on Principal Component analysis and median sampling. This approach avoids the pitfalls due to outliers and is both simple to implement and use across various incremental machine learning models. It also has independent usage as a sampling algorithm. We achieve better performance compared to state-of-the-art methods.

****

3D-IntPhys: Learning 3D Visual Intuitive Physics for Fluids, Rigid Bodies, and Granular Materials

Haotian Xue,Antonio Torralba,Daniel LK Yamins,Joshua B. Tenenbaum,Yunzhu Li,Hsiao-Yu Tung

Given a visual scene, humans have strong intuitions about how a scene can evolve over time under given actions. The intuition, often termed visual intuitive physics, is a critical ability that allows us to make effective plans to manipulate the scene to achieve desired outcomes without relying on extensive trial and error. In this paper, we present a framework capable of learning 3D-grounded visual intuitive physics models purely from unlabeled images. Our method is composed of a conditional Neural Radiance Field (NeRF)-style visual frontend and a 3D point-based dynamics prediction backend, in which we impose strong relational and structural inductive bias to capture the structure of the underlying environment. Unlike existing intuitive point-based dynamics works that rely on the supervision of dense point trajectory from simulators, we relax the requirements and only assume access to multi-view RGB images and (imperfect) instance masks. This enables the proposed model to handle scenarios where accurate point estimation and tracking are hard or impossible. We evaluate the models on three challenging scenarios involving fluid, granular materials, and rigid objects, where standard detection and tracking methods are not applicable. We show our model can make long-horizon future predictions by learning from raw images and significantly outperforms models that do not employ an explicit 3D representation space. We also show that, once trained, our model can achieve strong generalization in complex scenarios under extrapolate settings.

****

Continual Learning of Language Models

Zixuan Ke,Yijia Shao,Haowei Lin,Tatsuya Konishi,Gyuhak Kim,Bing Liu

Language models (LMs) have been instrumental for the rapid advance of natural language processing. This paper studies continual learning of LMs, in particular, continual domain-adaptive pre-training (or continual DAP-training). Existing research has shown that further pre-training an LM using a domain corpus to adapt the LM to the domain can improve the end-task performance in the domain. This paper proposes a novel method to continually DAP-train an LM with a sequence of unlabeled domain corpora to adapt the LM to these domains to improve their endtask performances. The key novelty of our method is a soft-masking mechanism that directly controls the update to the LM. A novel proxy is also proposed to preserve the general knowledge in the original LM. Additionally, it contrasts the representations of the previously learned domain knowledge (including the general knowledge in the pre-trained LM) and the knowledge from the current full network to a

chieve knowledge integration. The method not only overcomes catastrophic forgetting, but also achieves knowledge transfer to improve end-task performances. Empirical evaluation demonstrates the effectiveness of the proposed method.
****

Min-Max Multi-objective Bilevel Optimization with Applications in Robust Machine Learning

Alex Gu,Songtao Lu,Parikshit Ram,Tsui-Wei Weng

We consider a generic min-max multi-objective bilevel optimization problem with applications in robust machine learning such as representation learning and hyperparameter optimization. We design MORBiT, a novel single-loop gradient descent-ascent bilevel optimization algorithm, to solve the generic problem and present a novel analysis showing that MORBiT converges to the first-order stationary point at a rate of $\widetilde{\mathcal{O}}(n^{1/2} K^{-2/5})$ for a class of weakly convex problems with n objectives upon K iterations of the algorithm. Our analysis utilizes novel results to handle the non-smooth min-max multi-objective setup and to obtain a sublinear dependence in the number of objectives n. Experimental results on robust representation learning and robust hyperparameter optimization showcase (i) the advantages of considering the min-max multi-objective setup, and (ii) convergence properties of the proposed \morbit.
****

The Plug and Play of Language Models for Text-to-image Generation

Can Qin,Ning Yu,Chen Xing,Shu Zhang,Stefano Ermon,Yun Fu,Caiming Xiong,Ran Xu

Text-to-image (T2I) models enable controllable image generation through user-provided captions. A text encoder is typically used to map captions to a latent space, and it has been shown to be critical for model's performance. However, replacing or upgrading the text encoder in a T2I model is challenging due to the tight bond between the current encoder and the image decoder. It requires training the model from scratch, which can be prohibitively expensive. To address this problem, we introduce a more efficient approach to align a pre-trained language model with the latent space of an existing T2I model. We propose a Model Translation Network (MTN) and a new training objective to align the representation spaces of the two text encoders using only a corpus of unlabeled text. We empirically find that MTN can be trained efficiently and can boost the performance of existing T2I models by upgrading their text encoder. Moreover, we find that MTN can align multilingual language models such as XLM-Roberta, thus allowing existing T2I models to generate high-quality images from captions beyond English.
****

Learning Arborescence with An Efficient Inference Algorithm

Nan Jiang,Maxwell J Jacobson,Yexiang Xue

We consider a class of structured learning problems on arborescence (i.e., the directed spanning tree) from the input graph. The key step involved in this problem is predicting the minimal weight arborescence (MWA) from the learned model. In literature, there are two lines of research for predicting MWA: the Chu-Liu Edmonds (CLE) and the Lovasz methods. The CLE method is easy to implement while it takes $\mathcal{O}(n)$ cycle contractions. Here n is the graph size. The Lovasz method reduces to the multi-pair shortest path (MPSP) problem and takes only $\mathcal{O}(\log n)$ contractions. Nevertheless, in the CPU setting, MPSP has the same time complexity as finding MWA. The Lovasz method only attains time efficiency under a sufficient GPU setting. Both the aforementioned methods are painfully slow for large-scale learning tasks. In this research, we find the general MPSP problem can be simplified when working with machine learning models. This is because the learning model predicts edge weights for all pairs of vertices and the graph we process is always complete. Therefore, we only need to handle those paths that directly enter every weakly connected component (WCC) while the classic Lovasz method needs to handle all possible paths. This allows us to propose LAzy LoVAz (Lava) method that enjoys $\mathcal{O}(\log n)$ contractions as well as efficient performance in both CPU and GPU settings. In experiments, we consider synthetic datasets and two real-world learning tasks, i.e., graph-based dependency parsing and unsupervised parsing on ListOps. The empirical results exhibit important gains of our Lava method to the classic CLE and Lovasz methods,

that Lava boosts the training time for arborescence learning tasks.
****

Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning
Zeyuan Allen-Zhu,Yuanzhi Li
We formally study how \emph{ensemble} of deep learning models can improve test accuracy, and how the superior performance of ensemble can be distilled into a single model using \emph{knowledge distillation}. We consider the challenging case where the ensemble is simply an average of the outputs of a few independently trained neural networks with the \emph{same} architecture, trained using the \emph{same} algorithm on the \emph{same} data set, and they only differ by the random seeds used in the initialization.

We show that ensemble/knowledge distillation in \emph{deep learning} works very differently from traditional learning theory (such as boosting or NTKs). We develop a theory showing that when data has a structure we refer to as ``multi-view'', then ensemble of independently trained neural networks can provably improve test accuracy, and such superior test accuracy can also be provably distilled into a single model. Our result sheds light on how ensemble works in deep learning in a way that is completely different from traditional theorems, and how the ``dark knowledge'' is hidden in the outputs of the ensemble and can be used in distillation.
****

A Score-Based Model for Learning Neural Wavefunctions
Xuan Zhang,Shenglong Xu,Shuiwang Ji
Quantum Monte Carlo coupled with neural network wavefunctions has shown success in finding the ground state of quantum many-body systems. The existing optimization approaches compute the energy by sampling local energy from an explicit probability distribution given by the wavefunction. In this work, we provide a new optimization framework for obtaining properties of quantum many-body ground state using score-based neural networks. This new framework does not require explicit probability distribution and performs the sampling via Langevin dynamics. Our method is based on the key observation that the local energy is directly related to the score, defined as the gradient of the logarithmic wavefunction. Inspired by the score matching and the diffusion Monte Carlo methods, we derive a weighted score matching objective, which guides our score-based models to correctly converge to the ground state. We first validate our approach with experiments on quantum harmonic traps, and further results show that it can accurately learn the ground states of atomic systems. By implicitly modeling the high-dimensional data distribution, our work paves the way toward a more efficient representation of quantum systems.
****

Benchmarking Algorithms for Domain Generalization in Federated Learning
Ruqi Bai,Saurabh Bagchi,David I. Inouye
In this paper, we present a unified platform to study domain generalization in the federated learning (FL) context and conduct extensive empirical evaluations of the current state-of-the-art domain generalization algorithms adapted to FL. In particular, we perform a fair comparison of nine existing algorithms in solving domain generalization {either centralized domain generalization algorithms adapted to the FL context or existing FL domain generalization algorithms } to comprehensively explore the challenges introduced by FL. These challenges include statistical heterogeneity among clients, the number of clients, the number of communication rounds, etc. The evaluations are conducted on three diverse datasets including PACS (image dataset covering photo, sketch, cartoon, and painting domains), iWildCam (image dataset with 323 domains), and Py150 (natural language processing dataset with 8421 domains). The experiments show that the challenges brought by federated learning stay unsolved in the realistic experiment setting. Furthermore, the code base supports fair and reproducible new algorithm evaluation with easy implementation.
****

The Vendi Score: A Diversity Evaluation Metric for Machine Learning

Dan Friedman,Adji Bousso Dieng

Diversity is an important criterion for many areas of machine learning (ML), including generative modeling and dataset curation. Yet little work has gone into understanding, formalizing, and measuring diversity in ML. In this paper we address the diversity evaluation problem by proposing the Vendi Score, which connects and extends ideas from ecology and quantum statistical mechanics to ML. The Vendi Score is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix. This matrix is induced by a user-defined similarity function applied to the sample to be evaluated for diversity. In taking a similarity function as input, the Vendi Score enables its user to specify any desired form of diversity. Importantly, unlike many existing metrics in ML, the Vendi Score doesn’t require a reference dataset or distribution over samples or labels, it is therefore general and applicable to any generative model, decoding algorithm, and dataset from any domain where similarity can be defined. We showcase the Vendi Score on molecular generative modeling where we found it addresses shortcomings of the current diversity metric of choice in that domain. We also applied the Vendi Score to generative models of images and decoding algorithms of text where we found it confirms known results about diversity in those domains. Furthermore, we used the Vendi Score to measure mode collapse, a known shortcoming of generative adversarial networks (GANs). In particular, the Vendi Score revealed that even GANs that capture all the modes of a labelled dataset can be less diverse than the original dataset. Finally, the interpretability of the Vendi Score allowed us to diagnose several benchmark ML datasets for diversity, opening the door for diversity-informed data augmentation

****

Forward Super-Resolution: How Can GANs Learn Hierarchical Generative Models for Real-World Distributions

Zeyuan Allen-Zhu,Yuanzhi Li

Generative adversarial networks (GANs) are among the most successful models for learning high-complexity, real-world distributions. However, in theory, due to the highly non-convex, non-concave landscape of the minmax training objective, GAN remains one of the least understood deep learning models. In this work, we formally study how GANs can efficiently learn certain hierarchically generated distributions that are close to the distribution of real-life images. We prove that when a distribution has a structure that we refer to as \emph{forward super-resolution}, then simply training generative adversarial networks using stochastic gradient descent ascent (SGDA) can learn this distribution efficiently, both in sample and time complexities.

We also provide empirical evidence that our assumption ``forward super-resolution'' is very natural in practice, and the underlying learning mechanisms that we study in this paper (to allow us efficiently train GAN via GDA in theory) simulates the actual learning process of GANs on real-world problems.

****

Spotlight: Mobile UI Understanding using Vision-Language Models with a Focus

Gang Li,Yang Li

Mobile UI understanding is important for enabling various interaction tasks such as UI automation and accessibility. Previous mobile UI modeling often depends on the view hierarchy information of a screen, which directly provides the structural data of the UI, with the hope to bypass challenging tasks of visual modeling from screen pixels. However, view hierarchies are not always available, and are often corrupted with missing object descriptions or misaligned structure information. As a result, despite the use of view hierarchies could offer short-term gains, it may ultimately hinder the applicability and performance of the model. In this paper, we propose Spotlight, a vision-only approach for mobile UI understanding. Specifically, we enhance a vision-language model that only takes the screenshot of the UI and a region of interest on the screen---the focus---as the input. This general architecture of Spotlight is easily scalable and capable of performing a range of UI modeling tasks. Our experiments show that our model esta

blishes SoTA results on several representative UI tasks and outperforms previous methods that use both screenshots and view hierarchies as inputs. Furthermore, we explore multi-task learning and few-shot prompting capacities of the proposed models, demonstrating promising results in the multi-task learning direction.
****

A Control-Centric Benchmark for Video Prediction
Stephen Tian,Chelsea Finn,Jiajun Wu

Video is a promising source of knowledge for embodied agents to learn models of the world's dynamics. Large deep networks have become increasingly effective at modeling complex video data in a self-supervised manner, as evaluated by metrics based on human perceptual similarity or pixel-wise comparison. However, it remains unclear whether current metrics are accurate indicators of performance on downstream tasks. We find empirically that for planning robotic manipulation, existing metrics can be unreliable at predicting execution success. To address this, we propose a benchmark for action-conditioned video prediction in the form of a control benchmark that evaluates a given model for simulated robotic manipulation through sampling-based planning. Our benchmark, Video Prediction for Visual Planning (VP^2), includes simulated environments with 11 task categories and 310 task instance definitions, a full planning implementation, and training datasets containing scripted interaction trajectories for each task category. A central design goal of our benchmark is to expose a simple interface -- a single forward prediction call -- so it is straightforward to evaluate almost any action-conditioned video prediction model. We then leverage our benchmark to study the effects of scaling model size, quantity of training data, and model ensembling by analyzing five highly-performant video prediction models, finding that while scale can improve perceptual quality when modelling visually diverse settings, other attributes such as uncertainty awareness can also aid planning performance.
****

Continual Learning Based on Sub-Networks and Task Similarity
Zixuan Ke,Bing Liu,Wenhan Xiong,Asli Celikyilmaz,Haoran Li

Continual learning (CL) has two main objectives: preventing catastrophic forgetting (CF) and encouraging knowledge transfer (KT) across tasks. The existing literature mainly tries to overcome CF. Although some papers have focused on both CF and KT, they may still suffer from CF because of their ineffective handling of previous tasks and/or poor task similarity detection mechanisms to achieve KT. This work presents a new CL method that addresses the above issues. First, it overcomes CF by isolating the knowledge of each task via a learned mask that indicates a sub-network. Second, it proposes a novel technique to compute how important each mask is to the new task, which indicates how the new task is similar to an underlying old task. Similar tasks can share the same mask/subnetwork for KT, while dissimilar tasks use different masks/sub-networks for CF prevention. Comprehensive experiments have been conducted using a range of NLP problems, including classification, generation, and extraction to show that the proposed method consistently outperforms prior state-of-the-art baselines.
****

A Stable and Scalable Method for Solving Initial Value PDEs with Neural Networks
Marc Anton Finzi,Andres Potapczynski,Matthew Choptuik,Andrew Gordon Wilson

Unlike conventional grid and mesh based methods for solving partial differential equations (PDEs), neural networks have the potential to break the curse of dimensionality, providing approximate solutions to problems where using classical solvers is difficult or impossible. While global minimization of the PDE residual over the network parameters works well for boundary value problems, catastrophic forgetting impairs applicability to initial value problems (IVPs). In an alternative local-in-time approach, the optimization problem can be converted into an ordinary differential equation (ODE) on the network parameters and the solution propagated forward in time; however, we demonstrate that current methods based on this approach suffer from two key issues. First, following the ODE produces an uncontrolled growth in the conditioning of the problem, ultimately leading to unacceptably large numerical errors. Second, as the ODE methods scale cubically w ith the number of model parameters, they are restricted to small neural networks, significantly limiting their ability to represent intricate PDE initial conditions and solutions. Building on these insights, we develop Neural-IVP, an ODE based IVP solver which prevents the network from getting ill-conditioned and runs in time linear in the number of parameters, enabling us to evolve the dynamics of challenging PDEs with neural networks.

****

Shallow Learning In Materio.

Celestine Preetham Lawrence

We introduce Shallow Learning In Materio (SLIM) as a resource-efficient method to realize closed-loop higher-order perceptrons. Our SLIM method provides a rebuttal to the Minsky school's disputes with the Rosenblatt school about the efficacy of learning representations in shallow perceptrons. As a proof-of-concept, here we devise a physically-scalable realization of the parity function. Our findings are relevant to artificial intelligence engineers, as well as neuroscientists and biologists.

****

Data Subset Selection via Machine Teaching

Stephen Mussmann,Alex Fang,Ludwig Schmidt,Kevin Jamieson

We study the problem of data subset selection: given a fully labeled dataset and a training procedure, select a subset such that training on that subset yields approximately the same test performance as training on the full dataset. We propose an algorithm, inspired by recent work in machine teaching, that has theoretical guarantees, compelling empirical performance, and is model-agnostic meaning the algorithm's only information comes from the predictions of models trained on subsets. Furthermore, we prove lower bounds that show that our algorithm achieves a subset with near-optimal size (under computational hardness assumptions) while training on a number of subsets that is optimal up to extraneous log factors. We then empirically compare our algorithm, machine teaching algorithms, and coreset techniques on six common image datasets with convolutional neural networks. We find that our machine teaching algorithm can find a subset of CIFAR10 of size less than 16k that yields the same performance (5-6% error) as training on the full dataset of size 50k.

****

Do Summarization Models Synthesize?

Jay DeYoung,Iain James Marshall,Byron C Wallace

Multi-document summarization entails producing concise synopses of collections of inputs. For some applications, the synopsis should accurately \emph{synthesize} inputs with respect to a key property or aspect. For example, a synopsis of film reviews all written about a particular movie should reflect the average critic consensus. As a more consequential example, consider narrative summaries that accompany biomedical \emph{systematic reviews} of clinical trial results. These narratives should fairly summarize the potentially conflicting results from individual trials.

In this paper we ask: To what extent do modern multi-document summarization models implicitly perform this type of synthesis? To assess this we perform a suite of experiments that probe the degree to which conditional generation models trained for summarization using standard methods yield outputs that appropriately synthesize inputs. We find that existing models do partially perform synthesis, but do so imperfectly. In particular, they are over-sensitive to changes in input ordering and under-sensitive to changes in input compositions (e.g., the ratio of positive to negative movie reviews). We propose a simple, general method for improving model synthesis capabilities by generating an explicitly diverse set of candidate outputs, and then selecting from these the string best aligned with the expected aggregate measure for the inputs, or \emph{abstaining} when the model produces no good candidate. This approach improves model synthesis performance. Our hope is that by highlighting the need for synthesis (in some summarization settings), this work motivates further research into multi-document summarization methods and learning objectives that explicitly account for the need to synt

hesize.
****

CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets
Zachary Novack,Saurabh Garg,Zachary Chase Lipton
Open vocabulary models (e.g. CLIP) have shown strong performance on zeroshot classification through their ability generate embeddings for each class based on their (natural language) names. Prior work has focused on improving the accuracy of these models through prompt engineering or by incorporating a small amount of labeled downstream data (via finetuning). In this paper, we propose Classification with Hierarchical Label Sets (or CHiLS), an alternative strategy that proceeds in three steps: (i) for each class, produce a set of subclasses, using either existing label hierarchies or by querying GPT-3; (ii) perform the standard zero-shot CLIP procedure as though these subclasses were the labels of interest; (iii) map the predicted subclass back to its parent to produce the final prediction. Across numerous datasets, CHiLS leads to improved accuracy yielding gains of over 30% in situations where known hierarchies are available and more modest gains when they are not. CHiLS is simple to implement within existing CLIP pipelines and requires no additional training cost.
****

Multi-Grid Tensorized Fourier Neural Operator for High Resolution PDEs
Jean Kossaifi,Nikola Borislavov Kovachki,Kamyar Azizzadenesheli,Anima Anandkumar
Memory complexity and data scarcity are two main pressing challenges in learning solution operators of partial differential equations (PDE) at high resolutions. These challenges limited prior neural operator modelsMemory complexity and data scarcity are two main pressing challenges in learning solution operators of partial differential equations (PDE) at high resolutions. These challenges limited prior neural operator modelsMemory complexity and data scarcity are two main pressing challenges in learning solution operators of partial differential equations (PDE) at high resolutions. These challenges limited prior neural operator models to low/mid-resolution problems rather than full scale real-world problems. Yet, these problems possess spatially local structures that is not used by previous approaches. We propose to exploit this natural structure of real-world phenomena to predict solutions locally and unite them into a global solution. Specifically, we introduce a neural operator that scales to large resolutions by leveraging local and global structures through decomposition of both the input domain and the operator's parameter space. It consists of a multi-grid tensorized neural operator (MG-TFNO), a new data efficient and highly parallelizable operator learning approach with reduced memory requirement and better generalization. MG-TFNO employs a novel multi-grid based domain decomposition approach to exploit the spatially local structure in the data. Using the FNO as a backbone, its parameters are represented in a high-order latent subspace of the Fourier domain, through a global tensor factorization, resulting in an extreme reduction in the number of parameters and improved generalization. In addition, the low-rank regularization it applies to the parameters enables efficient learning in low-data regimes, which is particularly relevant for solving PDEs where obtaining ground-truth predictions is extremely costly and samples, therefore, are limited. We empirically verify the efficiency of our method on the turbulent Navier-Stokes equations where we demonstrate superior performance, with 2.5 times lower error, 10X compression of the model parameters, and 1.8X compression of the input domain size. Our tensorization approach yields up to 400x reduction in the number of parameter without loss in accuracy. Similarly, our domain decomposition method gives a 7x reduction in the domain size while slightly improving accuracy. Furthermore, our method can be trained with much fewer samples than previous approaches, outperforming the FNO when trained with just half the samples.
****

β-Stochastic Sign SGD: A Byzantine Resilient and Differentially Private Gradient Compressor for Federated Learning
Ming Xiang,Lili Su
Federated Learning (FL) is a nascent privacy-preserving learning framework under which the local data of participating clients is kept locally throughout model

training. Scarce communication resources and data heterogeneity are two defining characteristics of FL. Besides, a FL system is often implemented in a harsh environment -- leaving the clients vulnerable to Byzantine attacks. To the best of our knowledge, no gradient compressors simultaneously achieve quantitative Byzantine resilience and privacy preservation. In this paper, we fill this gap via revisiting the stochastic sign SGD \cite{Jin2020}. We propose β-stochastic sign SGD, which contains a gradient compressor that encodes a client's gradient information in sign bits subject to the privacy budget $\beta>0$. We show that as long as $\beta>0$, β-stochastic sign SGD converges in the presence of partial client participation and mobile Byzantine faults, showing that it achieves quantifiable Byzantine-resilience and differential privacy simultaneously. In sharp contrast, when $\beta=0$, the compressor is not differentially private. Notably, for the special case when each of the stochastic gradients involved is bounded with known bounds, our gradient compressor with $\beta=0$ coincides with the compressor proposed in \cite{Jin2020}. As a byproduct, we show that when the clients report sign messages, the popular information aggregation rules simple mean, trimmed mean, median and majority vote are identical in terms of the output signs. Our theories are corroborated by experiments on MNIST and CIFAR-10 datasets.

****

Sequential Brick Assembly with Efficient Constraint Satisfaction

Seokjun Ahn,Jungtaek Kim,Minsu Cho,Jaesik Park

We address the problem of generating a sequence of LEGO brick assembly with high-fidelity structures, satisfying physical constraints between bricks. The assembly problem is challenging since the number of possible structures increases exponentially with the number of available bricks, complicating the physical constraints to satisfy across bricks. To tackle this problem, our method performs a brick structure assessment to predict the next brick position and its confidence by employing a U-shaped sparse 3D convolutional network. The convolution filter efficiently validates physical constraints in a parallelizable and scalable manner, allowing to process of different brick types. To generate a novel structure, we devise a sampling strategy to determine the next brick position by considering attachable positions under physical constraints. Instead of using handcrafted brick assembly datasets, our model is trained with a large number of 3D objects that allow to create a new high-fidelity structure. We demonstrate that our method successfully generates diverse brick structures while handling two different brick types and outperforms existing methods based on Bayesian optimization, graph generative model, and reinforcement learning, all of which are limited to a single brick type.

****

Data-Efficient Finetuning Using Cross-Task Nearest Neighbors

Hamish Ivison,Noah A. Smith,Hannaneh Hajishirzi,Pradeep Dasigi

Language models trained on massive prompted multitask datasets like T0 (Sanh et al., 2021) or FLAN (Wei et al., 2021) can generalize to tasks unseen during training. We show that training on a carefully chosen subset of instances can outperform training on all available data on a variety of datasets. We assume access to a small number (250-1000) of unlabeled target task instances, select their nearest neighbors from a pool of multitask data, and use the retrieved data to train target task specific models. Our method is more data-efficient than training a single multitask model, while still outperforming it by large margins. We evaluate across a diverse set of tasks not in the multitask pool we retrieve from, including those used to evaluate T0 and in addition, more complex tasks including legal and scientific document QA. We retrieve small subsets of P3 (the collection of prompted datasets from which T0’s training data was sampled) and finetune T5 models that outperform the 3-billion parameter variant of T0 (T0-3B) by 8-30% on 11 out of 12 evaluation datasets while using at most 2% of the data used to train T0-3B. These models also provide a better initialization than T0-3B for few-shot finetuning on target-task data, as shown by a 3-23% relative improvement over few-shot finetuned T0-3B models on 8 datasets.

****

Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be

Frederik Kunstner,Jacques Chen,Jonathan Wilder Lavington,Mark Schmidt

The success of the Adam optimizer on a wide array of architectures has made it the default in settings where stochastic gradient descent (SGD) performs poorly. However, our theoretical understanding of this discrepancy is lagging, preventing the development of significant improvements on either algorithm. Recent work advances the hypothesis that Adam and other heuristics like gradient clipping outperform SGD on language tasks because the distribution of the error induced by sampling has heavy tails. This suggests that Adam outperform SGD because it uses a more robust gradient estimate. We evaluate this hypothesis by varying the batch size, up to the entire dataset, to control for stochasticity. We present evidence that stochasticity and heavy-tailed noise are not major factors in the performance gap between SGD and Adam. Rather, Adam performs better as the batch size increases, while SGD is less effective at taking advantage of the reduction in noise. This raises the question as to why Adam outperforms SGD in the full-batch setting. Through further investigation of simpler variants of SGD, we find that the behavior of Adam with large batches is similar to sign descent with momentum.

****

BiAdam: Fast Adaptive Bilevel Optimization Methods

Feihu Huang,Junyi Li,Shangqian Gao

Bilevel optimization recently has attracted increased interest in machine learning due to its many applications such as hyper-parameter optimization and mate learning. Although many bilevel optimization methods recently have been proposed, these methods do not consider using adaptive learning rates. It is well known that adaptive learning rates can accelerate many optimization algorithms including (stochastic) gradient-based algorithms. To fill this gap, in the paper, we propose a novel fast adaptive bilevel framework for solving bilevel optimization problems that the outer problem is possibly nonconvex and the inner problem is strongly convex. Our framework uses unified adaptive matrices including many types of adaptive learning rates, and can flexibly use the momentum and variance reduced techniques. In particular, we provide a useful convergence analysis framework for the bilevel optimization. Specifically, we propose a fast single-loop adaptive bilevel optimization (BiAdam) algorithm based on the basic momentum technique, which achieves a sample complexity of $\tilde{O}(\epsilon^{-4})$ for finding an ϵ-stationary point. Meanwhile, we propose an accelerated version of BiAdam algorithm (VR-BiAdam) by using variance reduced technique, which reaches the best known sample complexity of $\tilde{O}(\epsilon^{-3})$ without relying on large batch-size. To the best of our knowledge, we first study the adaptive bilevel optimization methods with adaptive learning rates. Some experimental results on data hyper-cleaning and hyper-representation learning tasks demonstrate the efficiency of the proposed algorithms.

****

Building Normalizing Flows with Stochastic Interpolants

Michael Samuel Albergo,Eric Vanden-Eijnden

A generative model based on a continuous-time normalizing flow between any pair of base and target probability densities is proposed. The velocity field of this flow is inferred from the probability current of a time-dependent density that interpolates between the base and the target in finite time. Unlike conventional normalizing flow inference methods based the maximum likelihood principle, which require costly backpropagation through ODE solvers, our interpolant approach leads to a simple quadratic loss for the velocity itself which is expressed in terms of expectations that are readily amenable to empirical estimation. The flow can be used to generate samples from either the base or target, and to estimate the likelihood at any time along the interpolant. In addition, the flow can be optimized to minimize the path length of the interpolant density, thereby paving the way for building optimal transport maps. In situations where the base is a Gaussian density, we also show that the velocity of our normalizing flow can also be used to construct a diffusion model to sample the target as well as estimate its score. However, our approach shows that we can bypass this diffusion completely and work at the level of the probability flow with greater simplicity, opening an avenue for methods based solely on ordinary differential equations as an alternative to those based on stochastic differential equations. Benchmarking on density estimation tasks illustrates that the learned flow can match and surpass conventional continuous flows at a fraction of the cost, and compares well with diffusions on image generation on CIFAR-10 and ImageNet 32×32. The method scales ab-initio ODE flows to previously unreachable image resolutions, demonstrated up to 128×128.

****

Elicitation Inference Optimization for Multi-Principal-Agent Alignment

Andrew Konya,Yeping Lina Qiu,Michael P Varga,Aviv Ovadya

In multi-principal-agent alignment scenarios spanning governance, markets, diplomacy, and AI, it is infeasible to elicit every principal's view on all perspectives relevant to agent decisions. Elicitation inference optimization (EIO) aims to minimize the n elicitations needed to approximate N principal's views across K perspectives. In this work, we demonstrate an EIO approach where data efficiency (NK/n) increases with scale. We introduce STUMP: an elicitation inference model which integrates an LLM with a latent factor model to enable learning transfer across samples, contexts, and languages. Then, we characterize STUMP's performance on a set of elicitation primitives from which scalable elicitation (sampling) protocols can be constructed. Building from these results, we design and demonstrate two scalable elicitation protocols for STUMP where data efficiency grows boundlessly, scaling like $O(n)$ in the number of elicitations n. This makes it possible to obtain complex, high-dimensional preference signals spanning principal populations at any scale.

****

Dual Student Networks for Data-Free Model Stealing

James Beetham,Navid Kardan,Ajmal Saeed Mian,Mubarak Shah

Data-free model stealing aims to replicate a target model without direct access to either the training data or the target model. To accomplish this, existing methods use a generator to produce samples in order to train a student model to match the target model outputs. To this end, the two main challenges are estimating gradients of the target model without access to its parameters, and generating a diverse set of training samples that thoroughly explores the input space. We propose a Dual Student method where two students are symmetrically trained in order to provide the generator a criterion to generate samples that the two students disagree on. On one hand, disagreement on a sample implies at least one student has classified the sample incorrectly when compared to the target model. This incentive towards disagreement implicitly encourages the generator to explore more diverse regions of the input space. On the other hand, our method utilizes gradients of student models to indirectly estimate gradients of the target model. We show that this novel training objective for the generator network is equivalent to optimizing a lower bound on the generator's loss if we had access to the target model gradients. In other words, our method alters the standard data-free model stealing paradigm by substituting the target model with a separate student model, thereby creating a lower bound which can be directly optimized without additional target model queries or separate synthetic datasets. We show that our new optimization framework provides more accurate gradient estimation of the target model and better accuracies on benchmark classification datasets. Additionally, our approach balances improved query efficiency with training computation cost. Finally, we demonstrate that our method serves as a better proxy model for transfer-based adversarial attacks than existing data-free model stealing methods.

****

Augmentation Curriculum Learning For Generalization in RL

Dylan Yung,Andrew Szot,Prithvijit Chattopadhyay,Judy Hoffman,Zsolt Kira

Many Reinforcement Learning tasks rely solely on pixel-based observations of the environment. During deployment, these observations can fall victim to visual perturbations and distortions, causing the agent’s policy to significantly degra

de
in performance. This motivates the need for robust agents that can generalize in the face of visual distribution shift. One common technique for doing this is to apply augmentations during training; however, it comes at the cost of performance. We propose Augmentation Curriculum Learning a novel curriculum learning approach that schedules augmentation into training into a weak augmentation phase and strong augmentation phase. We also introduce a novel visual augmentation strategy that proves to aid in the benchmarks we evaluate on. Our method achieves state-of-the-art performance on Deep Mind Control Generalization Benchmark.
****

Composite Slice Transformer: An Efficient Transformer with Composition of Multi-Scale Multi-Range Attentions
Mingu Lee,Saurabh Pitre,Tianyu Jiang,Pierre-David Letourneau,Matthew J Morse,Kanghwan Jang,Joseph Soriaga,Parham Noorzad,Hsin-Pai Cheng,Christopher Lott
Since the introduction of Transformers, researchers have tackled the notoriously expensive quadratic complexity problem. While significant computational efficiency improvements have been achieved, they come at the cost of reduced accuracy trade-offs. In this paper, we propose Composite Slice Transformer (CST), a Transformer-based network equipped with a composition of multi-scale multi-range attentions, boosting both efficiency and modeling capability.
After stacking fixed-length slices of the input sequence, each layer in CST performs a pair of fine-and-coarse-grained attentions with short-long ranges in a sequential manner, coupled with volatile instant positional embedding, enabling efficient token interactions {\em and} improving expressiveness of the model.
In addition to significantly reduced $O(NL+N^2/L^2)$ complexity for sequence length N and slice length L, CST achieves superior performance on a variety of tasks. We show that CST surpasses recently published efficient Transformers on the Long Range Arena benchmark, demonstrating the bidirectional long-range dependency modeling capability of our model. It outperforms the standard Transformer by a margin of 6.9\% in average accuracy across the five classification tasks of the benchmark, while being of complexity comparable to other efficient transformers. Furthermore, on the word-level autoregressive language modeling task with the WikiText-103 dataset, CST performs competitively against the Transformer model with only 2\% gap in the test perplexity while outperforming other efficient Transformers.
****

Graph Fourier MMD for signals on data graphs
Sam Leone,Alexander Tong,Guillaume Huguet,Guy Wolf,Smita Krishnaswamy
While numerous methods have been proposed for computing distances between probability distributions in Euclidean space, relatively little attention has been given to computing such distances for distributions on graphs. However, there has been a marked increase in data that either lies on graph (such as protein interaction networks) or can be modeled as a graph (single cell data), particularly in the biomedical sciences. Thus, it becomes important to find ways to compare signals defined on such graphs. Here, we propose Graph Fourier MMD (GFMMD), a novel a distance between distributions, or non-negative signals on graphs. GFMMD is defined via an optimal witness function that is both smooth on the graph and maximizes difference in expectation between the pair of distributions on the graph. We find an analytical solution to this optimization problem as well as an embedding of distributions that results from this method. We also prove several properties of this method including scale invariance and applicability to disconnected graphs. We showcase it on graph benchmark datasets as well on single cell RNA-sequencing data analysis. In the latter, we use the GFMMD-based gene embeddings to find meaningful gene clusters. We also propose a novel type of score for gene selection called {\em gene localization score} which helps select genes for cellular state space characterization.
****

Equal Improvability: A New Fairness Notion Considering the Long-term Impact

Ozgur Guldogan,Yuchen Zeng,Jy-yong Sohn,Ramtin Pedarsani,Kangwook Lee

Devising a fair classifier that does not discriminate against different groups is an important problem in machine learning. Although researchers have proposed various ways of defining group fairness, most of them only focused on the immediate fairness, ignoring the long-term impact of a fair classifier under the dynamic scenario where each individual can improve its feature over time. Such dynamic scenarios happen in real world, e.g., college admission and credit loaning, where each rejected sample makes effort to change its features to get accepted afterwards. In this dynamic setting, the long-term fairness should equalize the samples’ feature distribution across different groups after the rejected samples make some effort to improve. In order to promote long-term fairness, we propose a new fairness notion called Equal Improvability (EI), which equalizes the potential acceptance rate of the rejected samples across different groups assuming a bounded level of effort will be spent by each rejected sample. We analyze the properties of EI and its connections with existing fairness notions. To find a classifier that satisfies the EI requirement, we propose and study three different approaches that solve EI regularized optimization problems. Through experiments on both synthetic and real datasets, we demonstrate that the proposed EI-regularized algorithms encourage us to find a fair classifier in terms of EI. Finally, we provide experimental results on dynamic scenarios which highlight the advantages of our EI metric in achieving the long-term fairness. Codes are available in anonymous GitHub repository.

****

Does progress on ImageNet transfer to real world datasets?

Alex Fang,Simon Kornblith,Ludwig Schmidt

Does progress on ImageNet transfer to real world datasets? We investigate this question by evaluating ImageNet pre-trained models with varying accuracy (57% - 83%) on six practical image classification datasets. In particular, we study datasets collected with the goal of solving real world tasks (e.g., classifying images from camera traps or satellites), as opposed to web-scraped benchmarks collected for comparing models. On multiple datasets, models with higher ImageNet accuracy do not consistently yield performance improvements. For certain tasks, interventions such as data augmentation improve performance even when architectures do not. We hope that future benchmarks will include more diverse datasets to encourage a more comprehensive approach to improving learning algorithms.

****

Competitive Physics Informed Networks

Qi Zeng,Yash Kothari,Spencer H Bryngelson,Florian Tobias Schaefer

Neural networks can be trained to solve partial differential equations (PDEs) by using the PDE residual as the loss function. This strategy is called "physics-informed neural networks" (PINNs), but it currently cannot produce high-accuracy solutions, typically attaining about 0.1% relative error. We present an adversarial approach that overcomes this limitation, which we call competitive PINNs (CPINNs). CPINNs train a discriminator that is rewarded for predicting mistakes the PINN makes. The discriminator and PINN participate in a zero-sum game with the exact PDE solution as an optimal strategy. This approach avoids squaring the large condition numbers of PDE discretizations, which is the likely reason for failures of previous attempts to decrease PINN errors even on benign problems. Numerical experiments on a Poisson problem show that CPINNs achieve errors four orders of magnitude smaller than the best-performing PINN. We observe relative errors on the order of single-precision accuracy, consistently decreasing with each epoch. To the authors' knowledge, this is the first time this level of accuracy and convergence behavior has been achieved. Additional experiments on the nonlinear Schr{\"o}dinger, Burgers', and Allen--Cahn equation show that the benefits of CPINNs are not limited to linear problems.

****

Decomposed Prompting: A Modular Approach for Solving Complex Tasks

Tushar Khot,Harsh Trivedi,Matthew Finlayson,Yao Fu,Kyle Richardson,Peter Clark,Ashish Sabharwal

Few-shot prompting is a surprisingly powerful way to use Large Language Models (

LLMs) to solve various tasks. However, this approach struggles as the task complexity increases or when the individual reasoning steps of the task themselves are hard to learn, especially when embedded in more complex tasks. To address this, we propose Decomposed Prompting, a new approach to solve complex tasks by decomposing them (via prompting) into simpler sub-tasks that can be delegated to a library of prompting-based LLMs dedicated to these sub-tasks. This modular structure allows each prompt to be optimized for its specific sub-task, further decomposed if necessary, and even easily replaced with more effective prompts, trained models, or symbolic functions if desired.

We show that the flexibility and modularity of Decomposed Prompting allows it to outperform prior work on few-shot prompting using GPT3. On symbolic reasoning tasks, we can further decompose sub-tasks that are hard for LLMs into even simpler solvable sub-tasks. When the complexity comes from the input length, we can recursively decompose the task into the same task but with smaller inputs. We also evaluate our approach on textual multi-step reasoning tasks: on long-context multi-hop QA task, we can more effectively teach the sub-tasks via our separate sub-tasks prompts; and on open-domain multi-hop QA, we can incorporate a symbolic information retrieval within our decomposition framework, leading to improved performance on both tasks. Datasets, Code and Prompts available at https://github.com/allenai/DecomP.

****

Designing and Using Goal-Conditioned Tools

Ziang Liu,Stephen Tian,Michelle Guo,Karen Liu,Jiajun Wu

When limited by their own morphologies, humans and some species of animals have the remarkable ability to use objects from the environment towards accomplishing otherwise impossible tasks. Embodied agents might similarly unlock a range of additional capabilities through tool use. Recent techniques for jointly optimizing morphology and control via deep learning output effective solutions for tasks such as designing locomotion agents. But while designing a single-goal morphology makes sense for locomotion, manipulation involves a wide variety of strategies depending on the task goals at hand. An agent must be capable of rapidly prototyping specialized tools for different goals. Therefore, we propose the idea of learning a designer policy, rather than a single design. A designer policy is conditioned on task goals, and outputs a design for a tool that helps solve the task. A design-agnostic controller policy can then perform manipulation using these tools. In this work, we introduce a reinforcement learning framework for learning these policies. Through simulated manipulation tasks, we show that this framework is more sample efficient than black-box optimization methods in multi-goal settings. It can also perform zero-shot interpolation or finetuning to tackle previously unseen goals. Finally, we demonstrate that our framework allows tradeoffs between the complexity of design and control policies when required by practical constraints.

****

Post-mortem on a deep learning contest: a Simpson’s paradox and the complementary roles of scale metrics versus shape metrics

Michael W. Mahoney,charles h martin

To understand better good generalization performance in state-of-the-art neural network (NN) models, and in particular the success of the AlphaHat metric based on Heavy-Tailed Self-Regularization (HT-SR) theory, we analyze of a corpus of models that was made publicly-available for a contest to predict the generalization accuracy of NNs. These models include a wide range of qualities and were trained with a range of architectures and regularization hyperparameters. We break AlphaHat into its two subcomponent metrics: a scale-based metric; and a shape-based metric. We identify what amounts to a Simpson’s paradox: where “scale” metrics (from traditional statistical learning theory) perform well in aggregate, but can perform poorly on subpartitions of the data of a given depth, when regularization hyperparameters are varied; and where “shape” metrics (from HT-SR theory) perform well on each subpartition of the data, when hyperparameters are varied for models of a given depth, but can perform poorly overall when models with varying depths are aggregated. Our results highlight the subtlety of comparing models when both architectures and hyperparameters are varied; the complementary role of implicit scale versus implicit shape parameters in understanding NN model quality; and the need to go beyond one-size-fits-all metrics based on upper bounds from generalization theory to describe the performance of NN models. Our results also clarify further why the AlphaHat metric from HT-SR theory works so well at predicting generalization across a broad range of CV and NLP models.

****

ProtFIM: Fill-in-Middle Protein Sequence Design via Protein Language Models

Youhan Lee,Hasun Yu

Following the investigation that protein sequence determines its structure and function, engineering protein sequences allows us to optimize the functions of proteins for specific purposes such as enhancement of catalytic activity or binding affinity maturation. In protein engineering, there are many cases where the amino acids in the middle of a protein sequence are changed while maintaining the remaining residues to avoid unwanted functional changes from remaining residues. However, existing research on protein sequence design via protein language models (PLMs) has focused on modifying suffix residues by prompting prefix residues to the model or mutating the overall sequence residues. This is unsuitable for scenarios where the residues located in the middle of the sequence are to be optimized. In this work, we suggest a PLM-based framework to solve the fill-in-middle (FIM) protein engineering tasks. To evaluate the performance of PLMs on the FIM tasks, we design a novel evaluation scheme where PLMs are tasked to generate new sequences while maintaining the secondary structures. Also, we propose a new PROTein language model specialized for the Fill-In-Middle task, ProtFIM. Experiments confirm that ProtFIM performs FIM engineering efficiently, especially for alpha-helix structures, and provides decent protein representations of sequence-function relationships. Finally, we demonstrate an artificial protein sequence design framework composed of ProtFIM and a high-quality structure predictor as a novel tool to optimize protein sequences.

****

Beyond Deep Learning: An Evolutionary Feature Engineering Approach to Tabular Data Classification

Hengzhe Zhang,Qi Chen,Aimin Zhou,bing xue,Yan Wang,mengjie zhang

In recent years, deep learning has achieved impressive performance in the computer vision and natural language processing domains. In the tabular data classification scenario, with the emergence of the transformer architecture, a number of algorithms have been reported to yield better results than conventional tree-based models. Most of these methods attribute the success of deep learning methods to the expressive feature construction capability of neural networks. Nonetheless, in real practice, manually designed high-order features with traditional machine learning methods are still widely used because neural-network-based features can be easy to over-fitting. In this paper, we propose an evolution-based feature engineering algorithm to imitate the manual feature construction process through trial and improvement. Importantly, the evolutionary method provides an opportunity to optimize cross-validation loss, where gradient methods fail to do so. On a large-scale classification benchmark of 119 datasets, the experimental results demonstrate that the proposed method outperforms existing fine-tuned state-of-the-art tree-based and deep-learning-based classification algorithms.

****

Proportional Multicalibration

William La Cava,Elle Lett,Guangya Wan

Multicalibration is a desirable fairness criteria that constrains calibration error among flexibly-defined groups in the data while maintaining overall calibration. However, when outcome probabilities are correlated with group membership, multicalibrated models can exhibit a higher percent calibration error among groups with lower base rates than groups with higher base rates. As a result, it remains possible for a decision-maker to learn to trust or distrust model predictions for specific groups. To alleviate this, we propose proportional multicalibration, a criteria that constrains the percent calibration error among groups and within prediction bins. We prove that satisfying proportional multicalibration bou nds a model's multicalibration as well its differential calibration, a stronger fairness criteria inspired by the fairness notion of sufficiency. We provide an efficient algorithm for post-processing risk prediction models for proportional multicalibration and evaluate it empirically. We conduct simulation studies and investigate a real-world application of PMC-postprocessing to prediction of emergency department patient admissions. We observe that proportional multicalibration is a promising criteria for controlling simultenous measures of calibration fairness of a model over intersectional groups with virtually no cost in terms of classification performance.

****

On The Impact of Machine Learning Randomness on Group Fairness

Prakhar Ganesh,Hongyan Chang,Martin Strobel,Reza Shokri

Statistical measures for group fairness in machine learning reflect the gap in performance of algorithms across different groups. These measures, however, exhibit a high variance, between different training instances, that makes them unreliable for empirical evaluation of fairness. What is the cause of this variance, and how can we reduce it? We investigate the impact of different sources of randomness in machine learning on group fairness. We show that the variance in group fairness measures is mainly due to the high volatility of the learning process on under-represented groups, which itself is largely caused by the stochasticity of data order during training. Based on these findings, we show how to manipulate group level accuracy (i.e. model fairness), with high efficiency and negligible impact on the overall predictive power of the model, by changing the data order.

****

Using the Training History to Detect and Prevent Overfitting in Deep Learning Models

Hao Li,Gopi Krishnan Rajbahadur,Dayi Lin,Cor-Paul Bezemer,Zhen Jiang

Overfitting of deep learning models on training data leads to poor generalizability on unseen data. Overfitting can be (1) prevented (e.g., using dropout or early stopping) or (2) detected in a trained model (e.g., using correlation-based methods). We propose a method that can both detect and prevent overfitting based on the training history (i.e., validation losses). Our method first trains a time series classifier on training histories of overfit models. This classifier is then used to detect if a trained model is overfit. In addition, our trained classifier can be used to prevent overfitting by identifying the optimal point to stop a model's training. We evaluate our method on its ability to identify and prevent overfitting in real-world samples (collected from papers published in the last 5 years at top AI venues). We compare our method against correlation-based detection methods and the most commonly used prevention method (i.e., early stopping). Our method achieves an F1 score of 0.91 which is at least 5% higher than the current best-performing non-intrusive overfitting detection method. In addition, our method can find the optimal stopping point and avoid overfitting at least 32% earlier than early stopping and achieve at least the same accuracy (often better) as early stopping.

****

Multi-scale Sinusoidal Embeddings Enable Learning on High Resolution Mass Spectrometry Data

Gennady Voronov

Small molecules in biological samples are studied to provide information about disease states, environmental toxins, natural product drug discovery, and many other applications. The primary window into the composition of small molecule mixtures is tandem mass spectrometry (MS2), which produces data that are of high sensitivity and part per million resolution. We adopt multi-scale sinusoidal embeddings of the mass data in MS2 designed to meet the challenge of learning from the full resolution of MS2 data. Using these embeddings, we provide a new state of the art model for spectral library search, the standard task for initial evaluation of MS2 data. We also investigate the task of chemical property prediction from MS2 data, that has natural applications in high-throughput MS2 experiments and show that an average R^2 of 80\% for novel compounds can be achieved acros s 10 chemical properties prioritized by medicinal chemists. We use dimensionality reduction techniques and experiments with different floating point resolutions to show the essential role multi-scale sinusoidal embeddings play in learning from MS2 data.

****

Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors

Sizhe Chen,Geng Yuan,Xinwen Cheng,Yifan Gong,Minghai Qin,Yanzhi Wang,Xiaolin Huang

As data becomes increasingly vital, a company would be very cautious about releasing data, because the competitors could use it to train high-performance models, thereby posing a tremendous threat to the company's commercial competence. To prevent training good models on the data, we could add imperceptible perturbations to it. Since such perturbations aim at hurting the entire training process, they should reflect the vulnerability of DNN training, rather than that of a single model. Based on this new idea, we seek perturbed examples that are always unrecognized (never correctly classified) in training. In this paper, we uncover them by model checkpoints' gradients, forming the proposed self-ensemble protection (SEP), which is very effective because (1) learning on examples ignored during normal training tends to yield DNNs ignoring normal examples; (2) checkpoints' cross-model gradients are close to orthogonal, meaning that they are as diverse as DNNs with different architectures. That is, our amazing performance of ensemble only requires the computation of training one model. By extensive experiments with 9 baselines on 3 datasets and 5 architectures, SEP is verified to be a new state-of-the-art, e.g., our small $\ell_\infty=2/255$ perturbations reduce the accuracy of a CIFAR-10 ResNet18 from 94.56% to 14.68%, compared to 41.35% by the best-known method. Code is available at https://github.com/Sizhe-Chen/SEP.

****

Efficient parametric approximations of neural net function space distance

Nikita Dhawan,Sicong Huang,Juhan Bae,Roger Baker Grosse

It is often useful to compactly summarize important properties of a training dataset so that they can be used later without storing and/or iterating over the entire dataset. We consider a specific case of this: approximating the function space distance (FSD) over the training set, i.e. the average distance between the outputs of two neural networks. We propose an efficient approximation to FSD for ReLU neural networks based on approximating the architecture as a linear network with stochastic gating. Despite requiring only one parameter per unit of the network, our approach outcompetes other parametric approximations with larger memory requirements. Applied to continual learning, our parametric approximation is competitive with state-of-the-art nonparametric approximations which require storing many training examples. Furthermore, we show its efficacy in influence function estimation, allowing influence functions to be accurately estimated without iterating over the full dataset.

****

Systematic Generalization and Emergent Structures in Transformers Trained on Structured Tasks

Yuxuan Li,James McClelland

Transformer networks have seen great success in natural language processing and machine vision, where task objectives such as next word prediction and image classification benefit from nuanced context sensitivity across high-dimensional inputs. However, there is an ongoing debate about how and when transformers can acquire highly structured behavior and achieve systematic generalization. Here, we explore how well a causal transformer can perform a set of algorithmic tasks, including copying, sorting, and hierarchical compositions of these operations. We demonstrate strong generalization to sequences longer than those used in training by replacing the standard positional encoding typically used in transformers with labels arbitrarily paired with items in the sequence. By finding the layer and head configuration sufficient to solve the task, then performing ablation experiments and representation analysis, we show that two-layer transformers learn generalizable solutions to multi-level problems and develop signs of systematic task decomposition. They also exploit shared computation across related tasks. T hese results provide key insights into how transformer models may be capable of decomposing complex decisions into reusable, multi-level policies in tasks requiring structured behavior.

****

Energy-Inspired Self-Supervised Pretraining for Vision Models

Ze Wang,Jiang Wang,Zicheng Liu,Qiang Qiu

Motivated by the fact that forward and backward passes of a deep network naturally form symmetric mappings between input and output representations, we introduce a simple yet effective self-supervised vision model pretraining framework inspired by energy-based models (EBMs). In the proposed framework, we model energy estimation and data restoration as the forward and backward passes of a single network without any auxiliary components, e.g., an extra decoder. For the forward pass, we fit a network to an energy function that assigns low energy scores to samples that belong to an unlabeled dataset, and high energy otherwise. For the backward pass, we restore data from corrupted versions iteratively using gradient-based optimization along the direction of energy minimization. In this way, we naturally fold the encoder-decoder architecture widely used in masked image modeling into the forward and backward passes of a single vision model. Our framework accepts a wide range of pretext tasks with different data corruption methods, and permits models to be pretrained from masked image modeling, patch sorting, and image restoration, including super-resolution, denoising, and colorization. We support our findings with extensive experiments, and show the proposed method delivers comparable and even better performance with remarkably fewer epochs of training compared to the state-of-the-art self-supervised vision model pretraining methods. Our findings shed light on further exploring self-supervised vision model pretraining and pretext tasks beyond masked image modeling.

****

Effectively Modeling Time Series with Simple Discrete State Spaces

Michael Zhang,Khaled Kamal Saab,Michael Poli,Tri Dao,Karan Goel,Christopher Re

Time series modeling is a well-established problem, which often requires that methods (1) expressively represent complicated dependencies, (2) forecast long horizons, and (3) efficiently train over long sequences. State-space models (SSMs) are classical models for time series, and prior works combine SSMs with deep learning layers for efficient sequence modeling. However, we find fundamental limitations with these prior approaches, proving their SSM representations cannot express autoregressive time series processes. We thus introduce SpaceTime, a new state-space time series architecture that improves all three criteria. For expressivity, we propose a new SSM parameterization based on the companion matrix---a canonical representation for discrete-time processes---which enables SpaceTime's SSM layers to learn desirable autoregressive processes. For long horizon forecasting, we introduce a "closed-loop" variation of the companion SSM, which enables SpaceTime to predict many future time-steps by generating its own layer-wise inputs. For efficient training and inference, we introduce an algorithm that reduces the memory and compute of a forward pass with the companion matrix. With sequence length ℓ and state-space size d, we go from $\tilde{O}(d \ell)$ naïvely to $\tilde{O}(d + \ell)$. In experiments, our contributions lead to state-of-the-art results on extensive and diverse benchmarks, with best or second-best AUROC on 6 / 7 ECG and speech time series classification, and best MSE on 14 / 16 Informer forecasting tasks. Furthermore, we find SpaceTime (1) fits AR(p) processes that prior deep SSMs fail on, (2) forecasts notably more accurately on longer horizons than prior state-of-the-art, and (3) speeds up training on real-world ETTh1 data by 73% and 80% relative wall-clock time over Transformers and LSTMs.

****

Forgetful causal masking makes causal language models better zero-shot learners

Hao Liu,Xinyang Geng,Lisa Lee,Igor Mordatch,Sergey Levine,Sharan Narang,Pieter Abbeel

Large language models (LLM) trained using the next-token-prediction objective, such as GPT3 and PaLM, have revolutionized natural language processing in recent years by showing impressive zero-shot and few-shot capabilities across a wide ra nge of tasks. In this work, we propose a simple technique that significantly boosts the performance of LLMs without adding computational cost. Our key observation is that, by performing the next token prediction task with randomly selected past tokens masked out, we can improve the quality of the learned representations for downstream language understanding tasks. We hypothesize that randomly masking past tokens prevents over-attending to recent tokens and encourages attention to tokens in the distant past. By randomly masking input tokens in the PaLM model, we show that we can significantly improve PaLM's zero-shot performance on the SuperGLUE benchmark from 55.7 to 59.2. Experimental results show that FCM also improves PaLM's zero- and few-shot performance on a diverse suite of tasks, including commonsense reasoning, natural language inference and cloze completion. Moreover, we show that our technique also helps representation learning, significantly improving PaLM's finetuning results on SuperGLUE.

****

When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?

Mert Yuksekgonul,Federico Bianchi,Pratyusha Kalluri,Dan Jurafsky,James Zou

Despite the success of large vision and language models (VLMs) in many downstream applications, it is unclear how well they encode the compositional relationships between objects and attributes. Here, we create the Attribution, Relation, and Order (ARO) benchmark to systematically evaluate the ability of VLMs to understand different types of relationships, attributes, and order information. ARO consists of \emph{Visual Genome Attribution}, to test the understanding of objects' properties; \emph{Visual Genome Relation}, to test for relational understanding; and \emph{COCO-Order \& Flickr30k-Order}, to test for order sensitivity in VLMs. ARO is orders of magnitude larger than previous benchmarks of compositionality, with more than 50,000 test cases. We present the settings where state-of-the-art VLMs behave like bags-of-words---i.e. when they have poor relational understanding, can blunder when linking objects to their attributes, and demonstrate a severe lack of order sensitivity. VLMs are predominantly trained and evaluated on large scale datasets with rich compositional structure in the images and captions. Yet, training on these datasets has not been enough to address the lack of compositional understanding, and evaluating on these datasets has failed to surface this deficiency. To understand why these limitations emerge and are not represented in the standard tests, we zoom into the evaluation and training procedures. We demonstrate that it is possible to perform well on image-text retrieval over existing datasets without using the composition and order information. This further motivates the value of using ARO to benchmark VLMs. Given that contrastive pretraining optimizes for retrieval on large datasets with similar shortcuts, we hypothesize that this can explain why the models do not need to learn to represent compositional information. This finding suggests a natural solution: composition-aware hard negative mining. We show that a simple-to-implement modification of contrastive learning significantly improves the performance on tasks requiring understanding of order and compositionality.

****

A Time Series is Worth 64 Words: Long-term Forecasting with Transformers

Yuqi Nie,Nam H Nguyen,Phanwadee Sinthong,Jayant Kalagnanam

We propose an efficient design of Transformer-based models for multivariate time series forecasting and self-supervised representation learning. It is based on two key components: (i) segmentation of time series into subseries-level patches which are served as input tokens to Transformer; (ii) channel-independence where each channel contains a single univariate time series that shares the same embedding and Transformer weights across all the series. Patching design naturally has three-fold benefit: local semantic information is retained in the embedding; computation and memory usage of the attention maps are quadratically reduced given the same look-back window; and the model can attend longer history. Our channel-independent patch time series Transformer (PatchTST) can improve the long-term forecasting accuracy significantly when compared with that of SOTA Transformer-based models. We also apply our model to self-supervised pre-training tasks an d attain excellent fine-tuning performance, which outperforms supervised training on large datasets. Transferring of masked pre-training performed on one dataset to other datasets also produces SOTA forecasting accuracy.

****

Fantastic Rewards and How to Tame Them: A Case Study on Reward Learning for Task-oriented Dialogue Systems

Yihao Feng,Shentao Yang,Shujian Zhang,Jianguo Zhang,Caiming Xiong,Mingyuan Zhou,Huan Wang

When learning task-oriented dialogue (ToD) agents, reinforcement learning (RL) techniques can naturally be utilized to train dialogue strategies to achieve user-specific goals. Prior works mainly focus on adopting advanced RL techniques to train the ToD agents, while the design of the reward function is not well studied. This paper aims at answering the question of how to efficiently learn and leverage a reward function for training end-to-end (E2E) ToD agents. Specifically, we introduce two generalized objectives for reward-function learning, inspired by the classical learning-to-rank literature. Further, we utilize the learned reward function to guide the training of the E2E ToD agent. With the proposed techniques, we achieve competitive results on the E2E response-generation task on the Multiwoz 2.0 dataset. Source code and checkpoints are publicly released at https://github.com/Shentao-YANG/Fantastic_Reward_ICLR2023.

****

Supervision Complexity and its Role in Knowledge Distillation

Hrayr Harutyunyan,Ankit Singh Rawat,Aditya Krishna Menon,Seungyeon Kim,Sanjiv Kumar

Despite the popularity and efficacy of knowledge distillation, there is limited understanding of why it helps. In order to study the generalization behavior of a distilled student, we propose a new theoretical framework that leverages supervision complexity: a measure of alignment between teacher-provided supervision and the student's neural tangent kernel. The framework highlights a delicate interplay among the teacher's accuracy, the student's margin with respect to the teacher predictions, and the complexity of the teacher predictions. Specifically, it provides a rigorous justification for the utility of various techniques that are prevalent in the context of distillation, such as early stopping and temperature scaling. Our analysis further suggests the use of online distillation, where a student receives increasingly more complex supervision from teachers in different stages of their training. We demonstrate efficacy of online distillation and validate the theoretical findings on a range of image classification benchmarks and model architectures.

****

GLINKX: A Scalable Unified Framework For Homophilous and Heterophilous Graphs

Marios Papachristou,Rishab Goel,Frank Portman,Matthew Miller,Rong Jin

In graph learning, there have been two predominant inductive biases regarding graph-inspired architectures: On the one hand, higher-order interactions and message passing work well on homophilous graphs and are leveraged by GCNs and GATs. Such architectures, however, cannot easily scale to large real-world graphs. On the other hand, shallow (or node-level) models using ego features and adjacency embeddings work well in heterophilous graphs. In this work, we propose a novel scalable shallow method -- GLINKX -- that can work both on homophilous and heterophilous graphs. GLINKX leverages (i) novel monophilous label propagations (ii) ego/node features, (iii) knowledge graph embeddings as positional embeddings, (iv) node-level training, and (v) low-dimensional message passing. Formally, we prove novel error bounds and justify the components of GLINKX. Experimentally, we show its effectiveness of it on several homophilous and heterophilous datasets.

****

Marich: A Query-efficient & Online Model Extraction Attack using Public Data

Pratik Karmakar,Debabrota Basu

In this paper, we study black-box model stealing attacks where the attacker is only able to query a machine learning model through publicly available APIs. Specifically, our aim is to design a black-box model stealing attack that uses a minimal number of queries to create an informative replica of the target model. Fir

st, we reduce this problem to an online variational optimization problem. At every step, the attacker solves this problem to select the most informative query that maximizes the entropy of the selected queries and simultaneously reduces the mismatch between the target and the stolen models. We propose an online and adaptive algorithm, Marich, that leverages active learning to select the queries. We instantiate efficiency of our attack against different models, including logistic regression, BERT and ResNet18, trained on different text and image datasets. Marich is able to steal a model that can achieve 70-96$\%$ of true model's accuracy using 0.8-10$\%$ samples from the attack datasets which are publicly available and different from the training datasets. Our stolen models also achieve 75-98$\%$ accuracy of membership inference and also show 70-90$\%$ agreement of membership inference with direct membership inference on the target models. Our experiments validate that Marich is query-efficient and capable of creating an informative replica of the target model.
****

Lovasz Theta Contrastive Learning
Georgios Smyrnis,Matt Jordan,Ananya Uppal,Giannis Daras,Alex Dimakis
We establish a connection between the Lovasz theta function of a graph and the widely used InfoNCE loss. We show that under certain conditions, the minima of the InfoNCE loss are related to minimizing the Lovasz theta function on the empty similarity graph between the samples. Building on this connection, we generalize contrastive learning on weighted similarity graphs between samples. Our Lovasz theta contrastive loss uses a weighted graph that can be learned to take into account similarities between our data. We evaluate our method on image classification tasks, demonstrating an improvement of 1% in the supervised case and up to 4% in the unsupervised case.
****

Transferable Unlearnable Examples
Jie Ren,Han Xu,Yuxuan Wan,Xingjun Ma,Lichao Sun,Jiliang Tang
With more people publishing their personal data online, unauthorized data usage has become a serious concern. The unlearnable examples strategies have been introduced to prevent third parties from training on the data without permission. They add perturbations to the users’ data before publishing, so as to make the models trained on the perturbed published dataset invalidated. These perturbations have been generated for a specific training setting and a target dataset. However, their unlearnable effects significantly decrease when used in other training settings or datasets. To tackle this issue, we propose a novel unlearnable strategy based on Class-wise Separability Discriminant (CSD), which boosts the transferability of the unlearnable perturbations by enhancing the linear separability. Extensive experiments demonstrate the transferability of the unlearnable examples crafted by our proposed method across training settings and datasets.
****

MUG: Interactive Multimodal Grounding on User Interfaces
Tao Li,Gang Li,Jingjie Zheng,Purple Wang,Yang Li
We present MUG, a novel interactive task for multimodal grounding where a user and an agent work collaboratively on an interface screen. Prior works modeled multimodal UI grounding in one round: the user gives a command and the agent responds to the command. Yet, in a realistic scenario, a user command can be ambiguous when the target action is inherently difficult to articulate in natural language. MUG allows multiple rounds of interactions such that upon seeing the agent responses, the user can give further commands for the agent to refine or even correct its actions. Such interaction is critical for improving grounding performances in real-world use cases. To investigate the problem, we create a new dataset that consists of 77,820 sequences of human user-agent interaction on mobile interfaces in which 20% involves multiple rounds of interactions. To establish our benchmark, we experiment with a range of modeling variants and evaluation strategies, including both offline and online evaluation—the online strategy consists of both human evaluation and automatic with simulators. Our experiments show that allowing iterative interaction significantly improves the absolute task completion by 18% over the entire test dataset and 31% over the challenging subset. Our

results lay the foundation for further investigation of the problem.
****

Tabular Deep Learning when $d \gg n$ by Using an Auxiliary Knowledge Graph
Camilo Ruiz,Hongyu Ren,Kexin Huang,Jure Leskovec

Machine learning models exhibit strong performance on datasets with abundant labeled samples. However, for tabular datasets with extremely high d-dimensional features but limited n samples (i.e. $d \gg n$), machine learning models struggle to achieve strong performance. Here, our key insight is that even in tabular datasets with limited labeled data, input features often represent real-world entities about which there is abundant prior information which can be structured as an auxiliary knowledge graph (KG). For example, in a tabular medical dataset where every input feature is the amount of a gene in a patient's tumor and the label is the patient's survival, there is an auxiliary knowledge graph connecting gene names with drug, disease, and human anatomy nodes. We therefore propose PLATO, a machine learning model for tabular data with $d \gg n$ and an auxiliary KG with input features as nodes. PLATO uses a multilayer perceptron (MLP) to predict the output labels from the tabular data and the auxiliary KG with two methodological components. First, PLATO predicts the parameters in the first layer of the MLP from the auxiliary KG. PLATO thereby reduces the number of trainable parameters in the MLP and integrates auxiliary information about the input features. Second, PLATO predicts different parameters in the first layer of the MLP for every input sample, thereby increasing the MLP’s representational capacity by allowing it to use different prior information for every input sample. Across 10 state-of-the-art baselines and 6 $d \gg n$ datasets, PLATO exceeds or matches the prior state-of-the-art, achieving performance improvements of up to 10.19%. Overall, PLATO uses an auxiliary KG about input features to enable tabular deep learning prediction when $d \gg n$.
****

Random Laplacian Features for Learning with Hyperbolic Space
Tao Yu,Christopher De Sa

Due to its geometric properties, hyperbolic space can support high-fidelity embeddings of tree- and graph-structured data, upon which various hyperbolic networks have been developed. Existing hyperbolic networks encode geometric priors not only for the input, but also at every layer of the network. This approach involves repeatedly mapping to and from hyperbolic space, which makes these networks complicated to implement, computationally expensive to scale, and numerically unstable to train. In this paper, we propose a simpler approach: learn a hyperbolic embedding of the input, then map once from it to Euclidean space using a mapping that encodes geometric priors by respecting the isometries of hyperbolic space, and finish with a standard Euclidean network. The key insight is to use a random feature mapping via the eigenfunctions of the Laplace operator, which we show can approximate any isometry-invariant kernel on hyperbolic space. Our method can be used together with any graph neural networks: using even a linear graph model yields significant improvements in both efficiency and performance over other hyperbolic baselines in both transductive and inductive tasks.
****

Replay Memory as An Empirical MDP: Combining Conservative Estimation with Experience Replay
Hongming Zhang,Chenjun Xiao,Han Wang,Jun Jin,bo xu,Martin Müller

Experience replay, which stores transitions in a replay memory for repeated use, plays an important role of improving sample efficiency in reinforcement learning. Existing techniques such as reweighted sampling, episodic learning and reverse sweep update further process the information in the replay memory to make experience replay more efficient. In this work, we further exploit the information in the replay memory by treating it as an empirical \emph{Replay Memory MDP (RM-MDP)}. By solving it with dynamic programming, we learn a conservative value estimate that \emph{only} considers transitions observed in the replay memory. Both value and policy regularizers based on this conservative estimate are developed and integrated with model-free learning algorithms. We design the metric \textit{memory density} to measure the quality of RM-MDP. Our empirical studies quantit

atively find a strong correlation between performance improvement and memory density. Our method combines \emph{Conservative Estimation with Experience Replay (CEER)}, improving sample efficiency by a large margin, especially when the memory density is high. Even when the memory density is low, such a conservative estimate can still help to avoid suicidal actions and thereby improve performance.
****

Neural Causal Models for Counterfactual Identification and Estimation
Kevin Muyuan Xia,Yushu Pan,Elias Bareinboim
Evaluating hypothetical statements about how the world would be had a different course of action been taken is arguably one key capability expected from modern AI systems. Counterfactual reasoning underpins discussions in fairness, the determination of blame and responsibility, credit assignment, and regret. In this paper, we study the evaluation of counterfactual statements through neural models. Specifically, we tackle two causal problems required to make such evaluations, i.e., counterfactual identification and estimation from an arbitrary combination of observational and experimental data. First, we show that neural causal models (NCMs) are expressive enough and encode the structural constraints necessary for performing counterfactual reasoning. Second, we develop an algorithm for simultaneously identifying and estimating counterfactual distributions. We show that this algorithm is sound and complete for deciding counterfactual identification in general settings. Third, considering the practical implications of these results, we introduce a new strategy for modeling NCMs using generative adversarial networks. Simulations corroborate with the proposed methodology.
****

Connecting representation and generation via masked vision-language transformer
Xinyang Geng,Lisa Lee,Igor Mordatch,Sergey Levine,Pieter Abbeel,Hao Liu
Recently, there has been great progress in the self-supervised pre-training of multimodal representation models that understand image and language jointly. One particularly popular application of such models is text-to-image generation, which is typically obtained via a two-stage process: in the first stage, a representation model is trained via self-supervised objectives; then in the second stage, a conditional generative decoder is trained on top of the representation to generate natural images. In this work, we aim at bringing representation learning and conditional generation together by unifying the two stages into a single model and training objective. We present UPGen, a unified pre-trained model for both representation learning and generation. UPGen is trained with a simple masked token prediction objective on a flexible mixture of image and language data. We use a pre-trained VQGAN image tokenizer to convert images into discrete tokens, then train a masked token prediction model on both paired image-text datasets and unpaired language datasets, using randomly sampled mask ratios. We show that this masked token prediction model can be directly used to generate images and language by iteratively re-masking and predicting the masked tokens. We demonstrate empirically that UPGen serves as both a good representation learning model and a generative model for both image and language.
****

Is margin all you need? An extensive empirical study of active learning on tabular data
Dara Bahri,Heinrich Jiang,Tal Schuster,Afshin Rostamizadeh
Given a labeled training set and a collection of unlabeled data, the goal of active learning (AL) is to identify the best unlabeled points to label. In this comprehensive study, we analyze the performance of a variety of AL algorithms on deep neural networks trained on 69 real-world tabular classification datasets from the OpenML-CC18 benchmark. We consider different data regimes and the effect of self-supervised model pre-training. Surprisingly, we find that the classical margin sampling technique matches or outperforms all others, including current state-of-art, in a wide range of experimental settings. To researchers, we hope to encourage rigorous benchmarking against margin, and to practitioners facing tabular data labeling constraints that hyper-parameter-free margin may often be all they need.
****

Momentum Stiefel Optimizer, with Applications to Suitably-Orthogonal Attention, and Optimal Transport

Lingkai Kong,Yuqing Wang,Molei Tao

The problem of optimization on Stiefel manifold, i.e., minimizing functions of (not necessarily square) matrices that satisfy orthogonality constraints, has been extensively studied. Yet, a new approach is proposed based on, for the first time, an interplay between thoughtfully designed continuous and discrete dynamics. It leads to a gradient-based optimizer with intrinsically added momentum. This method exactly preserves the manifold structure but does not require additional operation to keep momentum in the changing (co)tangent space, and thus has low computational cost and pleasant accuracy. Its generalization to adaptive learning rates is also demonstrated. Notable performances are observed in practical tasks. For instance, we found that placing orthogonal constraints on attention heads of trained-from-scratch Vision Transformer (Dosovitskiy et al., 2020) could markedly improve its performance, when our optimizer is used, and it is better that each head is made orthogonal within itself but not necessarily to other heads. This optimizer also makes the useful notion of Projection Robust Wasserstein Distance (Paty and Cuturi, 2019; Lin et al., 2020) for high-dim. optimal transport even more effective.

****

Target Conditioned Representation Independence (TCRI); from Domain-Invariant to Domain-General Representations

Olawale Elijah Salaudeen,Oluwasanmi O Koyejo

We propose a Target Conditioned Representation Independence (TCRI) objective for domain generalization. TCRI addresses the limitations of existing domain generalization methods due to incomplete constraints. Specifically, TCRI implements regularizers motivated by conditional independence constraints that are sufficient to strictly learn complete sets of invariant mechanisms, which we show are necessary and sufficient for domain generalization. Empirically, we show that TCRI is effective on both synthetic and real-world data. TCRI is competitive with baselines in average accuracy while outperforming them in worst-domain accuracy, indicating desired cross-domain stability.

****

Multi-Task Option Learning and Discovery for Stochastic Path Planning

Naman Shah,Siddharth Srivastava

This paper addresses the problem of reliably and efficiently solving broad classes of long-horizon stochastic path planning problems. Starting with a vanilla RL formulation with a stochastic dynamics simulator and an occupancy matrix of the environment, our approach computes useful options with policies as well as high-level paths that compose the discovered options.

Our main contributions are (1) data-driven methods for creating abstract states that serve as endpoints for helpful options, (2) methods for computing option policies using auto-generated option guides in the form of dense pseudo-reward functions, and (3) an overarching algorithm for composing the computed options. We show that this approach yields strong guarantees of executability and solvability: under fairly general conditions, the computed option guides lead to composable option policies and consequently ensure downward refinability. Empirical evaluation on a range of robots, environments, and tasks shows that this approach effectively transfers knowledge across related tasks and that it outperforms existing approaches by a significant margin.

****

MolEBM: Molecule Generation and Design by Latent Space Energy-Based Modeling

Deqian Kong,Bo Pang,Tian Han,Ying Nian Wu

Generation of molecules with desired chemical and biological properties such as high drug-likeness, high binding affinity to target proteins, is critical in drug discovery. In this paper, we propose a probabilistic generative model to capture the joint distribution of molecules and their properties. Our model assumes an energy-based model (EBM) in the latent space. Given the latent vector sampled from the latent space EBM, both molecules and molecular properties are conditionally sampled via a top-down molecule generator model and a property regression m odel respectively. The EBM in a low dimensional latent space allows our model to capture complex chemical rules implicitly but efficiently and effectively. Due to the joint modeling with chemical properties, molecule design can be conveniently and naturally achieved by conditional sampling from our learned model given desired properties, in both single-objective and multi-objective optimization settings. The latent space EBM, top-down molecule generator, and property regression model are learned jointly by maximum likelihood, while optimization of properties is accomplished by gradual shifting of the model distribution towards the region supported by molecules with high property values. Our experiments show that our model outperforms state-of-the-art models on various molecule design tasks.

****

Information-Theoretic Diffusion

Xianghao Kong,Rob Brekelmans,Greg Ver Steeg

Denoising diffusion models have spurred significant gains in density modeling and image generation, precipitating an industrial revolution in text-guided AI art generation. We introduce a new mathematical foundation for diffusion models inspired by classic results in information theory that connect Information with Minimum Mean Square Error regression, the so-called I-MMSE relations. We generalize the I-MMSE relations to \emph{exactly} relate the data distribution to an optimal denoising regression problem, leading to an elegant refinement of existing diffusion bounds. This new insight leads to several improvements for probability distribution estimation, including a theoretical justification for diffusion model ensembling. Remarkably, our framework shows how continuous and discrete probabilities can be learned with the same regression objective, avoiding domain-specific generative models used in variational methods.

****

Bandwith Enables Generalization in Quantum Kernel Models

Abdulkadir Canatar,Evan Peters,Cengiz Pehlevan,Stefan M. Wild,Ruslan Shaydulin

Quantum computers are known to provide speedups over classical state-of-the-art machine learning methods in some specialized settings. For example, quantum kernel methods have been shown to provide an exponential speedup on a learning version of the discrete logarithm problem. Understanding the generalization of quantum models is essential to realizing similar speedups on practically interesting problems. Recent results demonstrate that generalization is hindered by the exponential size of the quantum feature space. Although these results suggest that quantum models cannot generalize when the number of qubits is large, in this paper we show that these results rely on overly restrictive assumptions. We consider a wider class of models by varying a hyperparameter that we call quantum kernel bandwidth. We analyze the large-qubit limit and provide explicit formulas for the generalization of a quantum model that can be solved in closed form. Specifically, we show that changing the value of bandwidth can take a model from provably not being able to generalize on any target function to good generalization for well-aligned targets. Our analysis shows how the bandwidth controls the spectrum of the kernel integral operator, and thereby the inductive bias of the model. We demonstrate empirically that our theory correctly predicts how varying the bandwidth affects generalization of quantum models on challenging datasets, including those far outside our theoretical assumptions. We discuss the implications of our results for quantum advantage in machine learning.

****

SpENCNN: Orchestrating Encoding and Sparsity for Fast Homomorphically Encrypted Neural Network Inference

Ran Ran,Xinwei Luo,Wei Wang,Tao Liu,Gang Quan,Wujie Wen

Homomorphic Encryption (HE) is a promising technology for protecting user's data privacy for Machine Learning as a Service (MLaaS) on public clouds. However, the computation overheads associated with the HE operations, which can be orders of magnitude slower than their counterparts for plaintexts, can lead to extremely high latency in neural network inference, seriously hindering its application in practice. While extensive neural network optimization techniques have been proposed, such as sparsification and pruning for plaintext domain, they cannot addr

ess this problem effectively. In this paper, we propose an HE-based CNN inference framework, i.e., SpENCNN, that can effectively exploit the single-instruction-multiple-data (SIMD) feature of the HE scheme to improve the CNN inference latency. In particular, we first develop a HE-group convolution technique that can partition channels among different groups based on the data size and ciphertext size, and then encode them into the same ciphertext in an interleaved manner, so as to dramatically reduce the bottlenecked operations in HE convolution. We further develop a sub-block weight pruning technique that can reduce more costly HE-operations for CNN convolutions. Our experiment results show that the SpENCNN-optimized CNN models can achieve overall speedups of 8.37x, 12.11x, and 19.26x for LeNet, VGG-5, and HEFNet, respectively, with negligible accuracy loss.
****

No Pairs Left Behind: Improving Metric Learning with Regularized Triplet Objective

A. Ali Heydari,Naghmeh Rezaei,Daniel McDuff,Javier L Prieto

We propose a novel formulation of the triplet objective function that improves metric learning without additional sample mining or overhead costs. Our approach aims to explicitly regularize the distance between the positive and negative samples in a triplet with respect to the anchor-negative distance. As an initial validation, we show that our method (called No Pairs Left Behind [NPLB]) improves upon the traditional and current state-of-the-art triplet objective formulations on standard benchmark datasets. To show the effectiveness and potentials of NPLB on real-world complex data, we evaluate our approach on a large-scale healthcare dataset (UK Biobank), demonstrating that the embeddings learned by our model significantly outperform all other current representations on tested downstream tasks. Additionally, we provide a new model-agnostic single-time health risk definition that, when used in tandem with the learned representations, achieves the most accurate prediction of a patient's future health complications. Our results indicate that NPLB is a simple, yet effective framework for improving existing deep metric learning models, showcasing the potential implications of deep metric learning in more complex applications, especially in the biological and healthcare domains.
****

Minimal Value-Equivalent Partial Models for Scalable and Robust Planning in Lifelong Reinforcement Learning

Safa Alver,Doina Precup

Learning models of the environment from pure interaction is often considered an essential component of building lifelong reinforcement learning agents. However, the common practice in model-based reinforcement learning is to learn models that model every aspect of the agent's environment, regardless of whether they are important in coming up with optimal decisions or not. In this paper, we argue that such models are not particularly well-suited for performing scalable and robust planning in lifelong reinforcement learning scenarios and we propose new kinds of models that only model the relevant aspects of the environment, which we call minimal value-equivalent partial models. After providing the formal definitions of these models, we provide theoretical results demonstrating the scalability advantages of performing planning with minimal value-equivalent partial models and then perform experiments to empirically illustrate our theoretical results. Finally, we provide some useful heuristics on how to learn such models with deep learning architectures and empirically demonstrate that models learned in such a way can allow for performing planning that is robust to distribution shifts and compounding model errors. Overall, both our theoretical and empirical results suggest that minimal value-equivalent partial models can provide significant benefits to performing scalable and robust planning in lifelong reinforcement learning scenarios.
****

Predictive Coding with Approximate Laplace Monte Carlo

Umais Zahid,Qinghai Guo,Karl Friston,Zafeirios Fountas

Predictive coding (PC) accounts of perception now form one of the dominant computational theories of the brain. Despite this, they have enjoyed little export to

the broader field of machine learning, where comparative generative models have flourished. In part, this has been due to the poor performance of models trained with standard implementations of PC when evaluated by both sample quality and marginal likelihood. By adopting the perspective of PC as a variational Bayes algorithm under the Laplace approximation, we identify the source of these deficits to lie in the exclusion of an associated Hessian term in the standard PC objective function. To remedy this, we make three primary contributions: we begin by suggesting a simple Monte Carlo estimated evidence lower bound which relies on sampling from the Hessian-parameterised variational posterior. We then derive a novel block diagonal approximation to the full Hessian matrix that has lower memory requirements and favourable mathematical properties. Lastly, we present an algorithm that combines our method with standard PC to reduce memory complexity further. We evaluate models trained with our approach against the standard PC framework on image benchmark datasets. Our approach produces higher log-likelihoods and qualitatively better samples that more closely capture the diversity of the data-generating distribution.
****

SIMPLE: A Gradient Estimator for k-Subset Sampling
kareem ahmed,Zhe Zeng,Mathias Niepert,Guy Van den Broeck

k-subset sampling is ubiquitous in machine learning, enabling regularization and interpretability through sparsity. The challenge lies in rendering k-subset sampling amenable to end-to-end learning. This has typically involved relaxing the reparameterized samples to allow for backpropagation, but introduces both bias and variance. In this work, we fall back to discrete k-subset sampling on the forward pass. This is coupled with using the gradient with respect to the exact marginals, computed efficiently, as a proxy for the true gradient. We show that our gradient estimator exhibits lower bias and variance compared to state-of-the-art estimators. Empirical results show improved performance on learning to explain and sparse models benchmarks. We provide an algorithm for computing the exact ELBO for the k-subset distribution, obtaining significantly lower loss compared to state-of-the-art discrete sparse VAEs. All of our algorithms are exact and efficient.
****

Transformers Implement First-Order Logic with Majority Quantifiers
William Merrill,Ashish Sabharwal

Characterizing the implicit structure of the computation within neural networks is a foundational problem in the area of deep learning interpretability. Can their inner decision process be captured symbolically in some familiar logic? We show that any transformer neural network can be translated into an equivalent fixed-size first-order logic formula which may also use majority quantifiers. The idea is to simulate transformers with highly uniform threshold circuits and leverage known theoretical connections between circuits and logic. Our findings also reveal the surprising fact that the entire transformer computation can be reduced merely to the division of two (large) integers. While our results are most pertinent for transformers, they apply equally to a broader class of neural network architectures, namely those with a fixed-depth uniform computation graph made up of standard neural net components, which includes feedforward and convolutional networks.
****

Robustness Evaluation Using Local Substitute Networks
Aleksei Kuvshinov,Crisitan Pavel,Stephan Günnemann

Robustness of a neural network against adversarial examples is an important topic when a deep classifier is applied in safety critical use cases like health care or autonomous driving. In order to assess the robustness, practitioners use a range of different tools ranging from the adversarial attacks to exact computation of the distance to the decision boundary. We use the fact that robustness of a neural network is a local property and empirically show that computing the same metrics for the smaller local substitute networks yields good estimates of the robustness for lower cost. To construct the substitute network we develop two pruning techniques that preserve the local properties of the initial network arou nd a given anchor point. Our experiments on CIFAR10 and MNIST datasets prove that this approach saves a significant amount of computing time and is especially beneficial for the larger models.

****

Learning Iterative Neural Optimizers for Image Steganography

Xiangyu Chen,Varsha Kishore,Kilian Q Weinberger

Image steganography is the process of concealing secret information in images through imperceptible changes.

Recent work has formulated this task as a classic constrained optimization problem. In this paper, we argue that image steganography is inherently performed on the (elusive) manifold of natural images, and propose an iterative neural network trained to perform the optimization steps. In contrast to classical optimization methods like L-BFGS or projected gradient descent, we train the neural network to also stay close to the manifold of natural images throughout the optimization. We show that our learned neural optimization is faster and more reliable than classical optimization approaches. In comparison to previous state-of-the-art encoder-decoder based steganography methods, it reduces the recovery error rate by multiple orders of magnitude and achieves zero error up to 3 bits per pixel (bpp) without the need for error-correcting codes.

****

Graph Neural Networks as Multi-View Learning

Haoyu Han,Xiaorui Liu,Haitao Mao,MohamadAli Torkamani,Feng Shi,Victor Lee,Jiliang Tang

Graph Neural Networks (GNNs) have demonstrated powerful representation capability in semi-supervised node classification. In this task, there are often three types of information -- graph structure, node features, and node labels. Existing GNNs usually leverage both node features and graph structure by feature transformation and aggregation, following end-to-end training via node labels. In this paper, we change our perspective by considering these three types of information as three views of nodes. This perspective motivates us to design a new GNN framework as multi-view learning which enables alternating optimization training instead of end-to-end training, resulting in significantly improved computation and memory efficiency. Extensive experiments with different settings demonstrate the effectiveness and efficiency of the proposed method.

****

Cramming: Training a language model on a single GPU in one day

Jonas Geiping,Tom Goldstein

Recent trends in language modeling have focused on increasing performance through scaling, and have resulted in an environment where training language models is out of reach for most researchers and practitioners. While most in the community are asking how to push the limits of extreme computation, we ask the opposite question:

How far can we get with a single GPU in just one day?

We investigate the downstream performance achievable with a transformer-based language model trained completely from scratch with masked language modeling for a single day on a single consumer GPU.

Aside from re-analyzing nearly all components of the pretraining pipeline for this scenario and providing a modified pipeline with performance close to BERT, we investigate why scaling down is hard, and which modifications actually improve performance in this scenario. We provide evidence that even in this constrained setting, performance closely follows scaling laws observed in large-compute settings. Through the lens of scaling laws, we categorize a range of recent improvements to training and architecture and discuss their merit and practical applicability (or lack thereof) for the limited compute setting.

****

BertNet: Harvesting Knowledge Graphs from Pretrained Language Models

Shibo Hao,Bowen Tan,Kaiwen Tang,Bin Ni,Hengzhe Zhang,Eric Xing,Zhiting Hu

Symbolic knowledge graphs (KGs) have been constructed either by expensive human crowdsourcing or with complex text mining pipelines. The emerging large pretrain ed language models (LMs), such as BERT, have shown to implicitly encode massive knowledge which can be queried with properly designed prompts. However, compared to the explicit KGs, the implict knowledge in the black-box LMs is often difficult to access or edit and lacks explainability. In this work, we aim at harvesting symbolic KGs from the LMs, and propose a new framework for automatic KG construction empowered by the neural LMs’ flexibility and scalability. Compared to prior works that often rely on large human annotated data or existing massive KGs, our approach requires only the minimal definition of relations as inputs, and hence is suitable for extracting knowledge of rich new relations that are instantly assigned and not available before. The framework automatically generates diverse prompts, and performs efficient knowledge search within a given LM for consistent outputs. The knowledge harvested with our approach shows competitive quality, diversity, and novelty. As a result, we derive from diverse LMs a family of new KGs (e.g., BERTNET and ROBERTANET) that contain a richer set of relations, including some complex ones (e.g., "A is capable of but not good at B") that cannot be extracted with previous methods. Besides, the resulting KGs also serve as a vehicle to interpret the respective source LMs, leading to new insights into the varying knowledge capability of different LMs.

****

How Hard is Trojan Detection in DNNs? Fooling Detectors With Evasive Trojans
Mantas Mazeika,Andy Zou,Akul Arora,Pavel Pleskov,Dawn Song,Dan Hendrycks,Bo Li,David Forsyth

As AI systems become more capable and widely used, a growing concern is the possibility for trojan attacks in which adversaries inject deep neural networks with hidden functionality. Recently, methods for detecting trojans have proven surprisingly effective against existing attacks. However, there is comparatively little work on whether trojans themselves could be rendered hard to detect. To fill this gap, we develop a general method for making trojans more evasive based on several novel techniques and observations. Our method combines distribution-matching, specificity, and randomization to eliminate distinguishing features of trojaned networks. Importantly, our method can be applied to various existing trojan attacks and is detector-agnostic. In experiments, we find that our evasive trojans reduce the efficacy of a wide range of detectors across numerous evaluation settings while maintaining high attack success rates. Moreover, we find that evasive trojans are also harder to reverse-engineer, underscoring the importance of developing more robust monitoring mechanisms for neural networks and clarifying the offence-defense balance of trojan detection.

****

Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding
Kenton Lee,Mandar Joshi,Iulia Raluca Turc,Hexiang Hu,Fangyu Liu,Julian Martin Eisenschlos,Urvashi Khandelwal,Peter Shaw,Ming-Wei Chang,Kristina Toutanova

Visually-situated language is ubiquitous---sources range from textbooks with diagrams to web pages with images and tables, to mobile apps with buttons and forms. Perhaps due to this diversity, previous work has typically relied on domain-specific recipes with limited sharing of the underlying data, model architectures, and objectives. We present Pix2Struct, a pretrained image-to-text model for purely visual language understanding, which can be finetuned on tasks containing visually-situated language. Pix2Struct is pretrained by learning to parse masked screenshots of web pages into simplified HTML. The web, with its richness of visual elements cleanly reflected in the HTML structure, provides a large source of pretraining data well suited to the diversity of downstream tasks. Intuitively, this objective subsumes common pretraining signals such as OCR, language modeling, image captioning. In addition to the novel pretraining strategy, we introduce a variable-resolution input representation and a more flexible integration of language and vision inputs, where language prompts such as questions are rendered directly on top of the input image. For the first time, we show that a single pretrained model can achieve state-of-the-art results in six out of nine tasks across four domains: documents, illustrations, user interfaces, and natural images.

****

Confidence-Conditioned Value Functions for Offline Reinforcement Learning

Joey Hong,Aviral Kumar,Sergey Levine

Offline reinforcement learning (RL) promises the ability to learn effective policies solely using existing, static datasets, without any costly online interaction. To do so, offline RL methods must handle distributional shift between the dataset and the learned policy. The most common approach is to learn conservative, or lower-bound, value functions, which underestimate the return of OOD actions. However, such methods exhibit one notable drawback: policies optimized on such value functions can only behave according to a fixed, possibly suboptimal, degree of conservatism. However, this can be alleviated if we instead are able to learn policies for varying degrees of conservatism at training time and devise a method to dynamically choose one of them during evaluation. To do so, in this work, we propose learning value functions that additionally condition on the degree of conservatism, which we dub confidence-conditioned value functions. We derive a new form of a Bellman backup that simultaneously learns Q-values for any degree of confidence with high probability. By conditioning on confidence, our value functions enable adaptive strategies during online evaluation by controlling for confidence level using the history of observations thus far. This approach can be implemented in practice by conditioning the Q-function from existing conservative algorithms on the confidence. We theoretically show that our learned value functions produce conservative estimates of the true value at any desired confidence. Finally, we empirically show that our algorithm outperforms existing conservative offline RL algorithms on multiple discrete control domains.

****

Current Anomaly Detectors are Anomalous: On Semantic Treatment of OOD Inputs

Ramneet Kaur,Xiayan Ji,Souradeep Dutta,Yahan Yang,Michele Caprio,Elena Bernardis,Oleg Sokolsky,Insup Lee

Machine learning models have achieved impressive performance across different modalities. It is well known that these models are prone to making mistakes on out-of-distribution inputs. OOD detection has, therefore, gained a lot of attention recently. We observe that most existing detectors use the distribution estimated by the training dataset for OOD detection. This can be a serious impediment since faulty OOD detectors can potentially restrict utility of the model. Such detectors, tied to the bias in data collection process, can be impermeable to inputs lying outside the training distribution but with the same semantic information (e.g., class labels) as the training data. We argue that in-distribution should not be tied to just the training distribution but to the distribution of the semantic information contained in the training data. To support our argument, we perform OOD detection on semantic information extracted from the training data of MNIST and COCO datasets, and show that it not only reduces false alarms but also significantly improves detection of OOD inputs with spurious features from training data.

****

FedX: Federated Learning for Compositional Pairwise Risk Optimization

Zhishuai Guo,Rong Jin,Jiebo Luo,Tianbao Yang

In this paper, we tackle a novel federated learning (FL) problem for optimizing a family of compositional pairwise risks, to which no existing FL algorithms are applicable. In particular, the objective has the form of $\E_{\z\sim \mathcal S_1} f(\E_{\z'\sim\mathcal S_2} \ell(\w; \z, \z'))$, where two sets of data $\mathcal S_1, \mathcal S_2$ are distributed over multiple machines, $\ell(\cdot; \cdot,\cdot)$ is a pairwise loss that only depends on the prediction outputs of the input data pairs $(\z, \z')$, and $f(\cdot)$ is possibly a non-linear non-convex function. This problem has important applications in machine learning, e.g., AUROC maximization with a pairwise loss, and partial AUROC maximization with a compositional loss, etc. The challenges for designing a FL algorithm lie at the non-decomposability of the objective over multiple machines and the interdependency between different machines. We propose two provable FL algorithms (FedX) for handling linear and nolinear f, respectively. To tackle the challenges, we decouple the gradient's components with two types namely active parts and lazy part

s, where the {\it active} parts depend on local data that can be computed with the local model and the {\it lazy} parts depend on other machines that are communicated/computed based on historical models. We develop a novel theoretical analysis to address the issue of latency of lazy parts and interdependency between the local gradient estimators and the involved data. We establish both iteration and communication complexities and exhibit that using the historical models for computing the lazy parts do not degrade the complexity results. We conduct empirical studies of FedX for AUROC and partial AUROC maximization, and demonstrate their performance compared with multiple baselines.
****
On the Sensitivity of Reward Inference to Misspecified Human Models
Joey Hong,Kush Bhatia,Anca Dragan
Inferring reward functions from human behavior is at the center of value alignment – aligning AI objectives with what we, humans, actually want. But doing so relies on models of how humans behave given their objectives. After decades of research in cognitive science, neuroscience, and behavioral economics, obtaining accurate human models remains an open research topic. This begs the question: how accurate do these models need to be in order for the reward inference to be accurate? On the one hand, if small errors in the model can lead to catastrophic error in inference, the entire framework of reward learning seems ill-fated, as we will never have perfect models of human behavior. On the other hand, if as our models improve, we can have a guarantee that reward accuracy also improves, this would show the benefit of more work on the modeling side. We study this question both theoretically and empirically. We do show that it is unfortunately possible to construct small adversarial biases in behavior that lead to arbitrarily large errors in the inferred reward. However, and arguably more importantly, we are also able to identify reasonable assumptions under which the reward inference error can be bounded linearly in the error in the human model. Finally, we verify our theoretical insights in discrete and continuous control tasks with simulated and human data.
****
DeepDFA: Dataflow Analysis-Guided Efficient Graph Learning for Vulnerability Detection
Benjamin Steenhoek,Wei Le,Hongyang Gao
Deep learning-based vulnerability detection models have recently been shown to be effective and, in some cases, outperform static analysis tools. However, the highest-performing approaches use token-based transformer models, which do not leverage domain knowledge. Classical program analysis techniques such as dataflow analysis can detect many types of bugs and are the most commonly used methods in practice. Motivated by the causal relationship between bugs and dataflow analysis, we present DeepDFA, a dataflow analysis-guided graph learning framework and embedding that use program semantic features for vulnerability detection. We show that DeepDFA is performant and efficient. DeepDFA ranked first in recall, first in generalizing over unseen projects, and second in F1 among all the state-of-the-art models we experimented with. It is also the smallest model in terms of the number of parameters, and was trained in 9 minutes, 69x faster than the highest-performing baseline. DeepDFA can be used with other models. By integrating LineVul and DeepDFA, we achieved the best vulnerability detection performance of 96.4 F1 score, 98.69 precision, and 94.22 recall.
****
Probability flow solution of the Fokker-Planck equation
Nicholas Matthew Boffi,Eric Vanden-Eijnden
The method of choice for integrating the time-dependent Fokker-Planck equation in high-dimension is to generate samples from the solution via integration of the associated stochastic differential equation. Here, we introduce an alternative scheme based on integrating an ordinary differential equation that describes the flow of probability. Acting as a transport map, this equation deterministically pushes samples from the initial density onto samples from the solution at any later time. Unlike integration of the stochastic dynamics, the method has the advantage of giving direct access to quantities that are challenging to estimate fr

om trajectories alone, such as the probability current, the density itself, and its entropy. The probability flow equation depends on the gradient of the logarithm of the solution (its "score"), and so is a-priori unknown. To resolve this dependence, we model the score with a deep neural network that is learned on-the-fly by propagating a set of samples according to the instantaneous probability current. We consider several high-dimensional examples from the physics of interacting particle systems to highlight the efficiency and precision of the approach; we find that the method accurately matches analytical solutions computed by hand and moments computed via Monte-Carlo.

****

Binding Language Models in Symbolic Languages

Zhoujun Cheng,Tianbao Xie,Peng Shi,Chengzu Li,Rahul Nadkarni,Yushi Hu,Caiming Xiong,Dragomir Radev,Mari Ostendorf,Luke Zettlemoyer,Noah A. Smith,Tao Yu

Though end-to-end neural approaches have recently been dominating NLP tasks in both performance and ease-of-use, they lack interpretability and robustness. We propose Binder, a training-free neural-symbolic framework that maps the task input to a program, which (1) allows binding a unified API of language model (LM) functionalities to a programming language (e.g., SQL, Python) to extend its grammar coverage and thus tackle more diverse questions, (2) adopts an LM as both the program parser and the underlying model called by the API during execution, and (3) requires only a few in-context exemplar annotations. Specifically, we employ GPT-3 Codex as the LM. In the parsing stage, with only a few in-context exemplars, Codex is able to identify the part of the task input that cannot be answerable by the original programming language, correctly generate API calls to prompt Codex to solve the unanswerable part, and identify where to place the API calls while being compatible with the original grammar. In the execution stage, Codex can perform versatile functionalities (e.g., commonsense QA, information extraction) given proper prompts in the API calls. Binder achieves state-of-the-art results on WikiTableQuestions and TabFact datasets, with explicit output programs that benefit human debugging. Note that previous best systems are all finetuned on tens of thousands of task-specific samples, while Binder only uses dozens of annotations as in-context exemplars without any training. Our code is available at anonymized.

****

Probabilistic Categorical Adversarial Attack and Adversarial Training

Pengfei He,Han Xu,Jie Ren,Yuxuan Wan,Zitao Liu,Jiliang Tang

The existence of adversarial examples brings huge concern for people to apply Deep Neural Networks (DNNs) in safety-critical tasks. However, how to generate adversarial examples with categorical data is an important problem but lacks extensive exploration. Previously established methods leverage greedy search methods, which can be very time-consuming to conduct a successful attack. This also limits the development of adversarial training and potential defenses for categorical data. To tackle this problem, we propose a Probabilistic Categorical Adversarial Attack (PCAA), which transfers the discrete optimization problem to a continuous problem that can be solved efficiently by Projected Gradient Descent. In our paper, we theoretically analyze its optimality and time complexity to demonstrate its significant advantage over current greedy-based attacks. Moreover, based on our attack, we propose an efficient adversarial training framework. Through a comprehensive empirical study, we justify the effectiveness of our proposed attack and defense algorithms.

****

Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection

Jinhyung Park,Chenfeng Xu,Shijia Yang,Kurt Keutzer,Kris M. Kitani,Masayoshi Tomizuka,Wei Zhan

While recent camera-only 3D detection methods leverage multiple timesteps, the limited history they use significantly hampers the extent to which temporal fusion can improve object perception. Observing that existing works' fusion of multi-frame images are instances of temporal stereo matching, we find that performance is hindered by the interplay between 1) the low granularity of matching resolut ion and 2) the sub-optimal multi-view setup produced by limited history usage. Our theoretical and empirical analysis demonstrates that the optimal temporal difference between views varies significantly for different pixels and depths, making it necessary to fuse many timesteps over long-term history. Building on our investigation, we propose to generate a cost volume from a long history of image observations, compensating for the coarse but efficient matching resolution with a more optimal multi-view matching setup. Further, we augment the per-frame monocular depth predictions used for long-term, coarse matching with short-term, fine-grained matching and find that long and short term temporal fusion are highly complementary. While maintaining high efficiency, our framework sets new state-of-the-art on nuScenes, achieving first place on the test set and outperforming previous best art by 5.2% mAP and 3.7% NDS on the validation set. Code will be released here: https://github.com/Divadi/SOLOFusion.

****

Greedy Information Maximization for Online Feature Selection

Ian Connick Covert,Wei Qiu,MingYu Lu,Na Yoon Kim,Su-In Lee

Feature selection is commonly used to reduce feature acquisition costs, but the standard approach is to train models with static feature subsets. Here, we consider the online feature selection problem, where the model can adaptively query features based on the presently available information. Online feature selection has mainly been viewed as a reinforcement learning problem, but we propose a simpler approach of greedily selecting features that maximize mutual information with the response variable. This intuitive idea is difficult to implement without perfect knowledge of the joint data distribution, so we propose a deep learning approach that recovers the greedy procedure when perfectly optimized. We apply our approach to numerous datasets and observe better performance than both RL-based and offline feature selection methods

****

Towards Fair Classification against Poisoning Attacks

Han Xu,Xiaorui Liu,Yuxuan Wan,Jiliang Tang

Fair classification aims to stress the classification models to achieve the equality (treatment or prediction quality) among different sensitive groups. However, fair classification can be under the risk of poisoning attacks which deliberately insert malicious training samples to manipulate the trained classifiers' performance. In this work, we study the poisoning scenario where the attacker can insert a small fraction of samples into training data, with arbitrary sensitive attributes as well as other predictive features. We demonstrate that the fairly trained classifiers can be greatly vulnerable to such poisoning attacks, with much worse accuracy & fairness trade-off, even when we apply some of the most effective defenses (originally proposed to defend traditional classification tasks). As countermeasures to defend fair classification tasks, we propose a general and theoretically guaranteed framework which accommodates traditional defense methods to fair classification against poisoning attacks. Through extensive experiments, the results validate that the proposed defense framework obtains better robustness in terms of accuracy and fairness than baseline methods.

****

Unveiling Transformers with LEGO: A Synthetic Reasoning Task

Yi Zhang,Arturs Backurs,Sebastien Bubeck,Ronen Eldan,Suriya Gunasekar,Tal Wagner

We propose a synthetic reasoning task, LEGO (Learning Equality and Group Operations), that encapsulates the problem of following a chain of reasoning, and we study how the Transformer architectures learn this task. We pay special attention to data effects such as pretraining (on seemingly unrelated NLP tasks) and dataset composition (e.g., differing chain length at training and test time), as well as architectural variants such as weight-tied layers or adding convolutional components. We study how the trained models eventually succeed at the task, and in particular, we are able to understand (to some extent) some of the attention heads as well as how the information flows in the network. Based on these observations we propose a hypothesis that here pretraining helps for LEGO tasks due to certain structured attention patterns, and we experimentally verify this hypothesis. We also observe that in some data regimes the trained transformer finds ``sh ortcut" solutions to follow the chain of reasoning, which impedes the model's robustness, and moreover we propose ways to prevent it. Motivated by our findings on structured attention patterns, we propose to replace certain attention heads with hardcoded patterns. This architectural change significantly reduces Flops and maintains or even improves the model's performance at large-scale pretraining.

****

How Much Data Are Augmentations Worth? An Investigation into Scaling Laws, Invariance, and Implicit Regularization

Jonas Geiping,Micah Goldblum,Gowthami Somepalli,Ravid Shwartz-Ziv,Tom Goldstein,Andrew Gordon Wilson

Despite the clear performance benefits of data augmentations, little is known about why they are so effective. In this paper, we disentangle several key mechanisms through which data augmentations operate. Establishing an exchange rate between augmented and additional real data, we find that in out-of-distribution testing scenarios, augmentations which yield samples that are diverse, but inconsistent with the data distribution can be even more valuable than additional training data. Moreover, we find that data augmentations which encourage invariances can be more valuable than invariance alone, especially on small and medium sized training sets. Following this observation, we show that augmentations induce additional stochasticity during training, effectively flattening the loss landscape.

****

Spatial Reasoning Network for Zero-shot Constrained Scene Generation

Maxwell J Jacobson,Yexiang Xue

Constrained scene generation (CSG) generates images satisfying a given set of constraints. Zero-shot CSG generates images satisfying constraints not presented in the training set without retraining. Recent neural-based models generate images with excellent details, but largely cannot satisfy constraints, especially in complex scenes involving multiple objects. Such difficulty is due to the lack of effective approaches combining low-level visual element generation with high-level spatial reasoning. We introduce a Spatial Reasoning Network for constrained scene generation (SPREN). SPREN adds to the state-of-the-art image generation networks (for low-level visual element generation) a spatial reasoning module (for high-level spatial reasoning). The spatial reasoning module decides objects' positions following the output of a Recursive Neural Network (RNN), which is trained to learn implicit spatial knowledge (such as trees growing from the ground) from an image dataset. During inference, explicit constraints can be enforced by a forward-checking algorithm, which blocks invalid decisions from the RNN in a zero-shot manner. In experiments, we demonstrate SPREN is able to generate images with excellent detail while satisfying complex spatial constraints. SPREN also transfers good quality scene generation to unseen constraints without retraining.

****

Robust Graph Dictionary Learning

Weijie Liu,Jiahao Xie,Chao Zhang,Makoto Yamada,Nenggan Zheng,Hui Qian

Traditional Dictionary Learning (DL) aims to approximate data vectors as sparse linear combinations of basis elements (atoms) and is widely used in machine learning, computer vision, and signal processing. To extend DL to graphs, Vincent-Cuaz et al. 2021 propose a method, called GDL, which describes the topology of each graph with a pairwise relation matrix (PRM) and compares PRMs via the Gromov-Wasserstein Discrepancy (GWD). However, the lack of robustness often excludes GDL from a variety of real-world applications since GWD is sensitive to the structural noise in graphs. This paper proposes an improved graph dictionary learning algorithm based on a robust Gromov-Wasserstein discrepancy (RGWD) which has theoretically sound properties and an efficient numerical scheme. Based on such a discrepancy, our dictionary learning algorithm can learn atoms from noisy graph data. Experimental results demonstrate that our algorithm achieves good performance on both simulated and real-world datasets.

****

Matrix factorization under the constraint of connectivity between observed and s ource data ~ Muscle synergy analysis based on connectivity between muscle and brain activities ~

Takashi Isezaki,Michiaki Suzuki,Yukio Koike,Ryosuke Aoki,Yukio Nishimura

Matrix factorization is a popular method to investigate the hidden elements in observed data for tasks such as speech separation and muscle synergy analysis. The hidden elements may be closely related to the source phenomenon that cause the observed phenomenon.

However, conventional methods do not always factorize the observed phenomenon elements with the connectivity between the observed and source phenomena because they only use the observed phenomenon. This paper proposes a matrix decomposition method that constrains the connectivity between observed and source data by using the representations from a decoding model from source data to observed data. We applied our method to the corticomuscular system, which is made up of corticospinal pathways between the primary motor cortex and muscles in the body and creates muscle synergies that enable efficient connections between the brain and muscles. In this context, muscle activities are the observed phenomenon and brain activities are the source. Many previous studies have analyzed muscle synergies using only observed muscle activity, but there may be unrevealed muscle synergies under the constraint of the connectivity between brain and muscle activity. We therefore simultaneously recorded the brain activity from multiple regions of an extensive cortical area and the activity of multiple muscles of a monkey's forelimb while it performed a reach and grasp task throughout the course of recovery from a partial spinal cord injury (SCI). Analysis from a dataset of the monkey before SCI showed that some of the muscle synergies calculated from the proposed method using brain and muscle activities, did not exhibit a high degree of similarity to synergies obtained from the conventional method. The proposed method results obtained from the monkey after SCI showed an adaptive change in the number of muscle synergies associated with the degree of functional recovery. Specifically, the numbers of muscle synergies obtained by the proposed method initially increased immediately after SCI and then gradually decreased, while those obtained by a conventional method maintained the same number before and after SCI. These results suggest that our method is able to capture the unrevealed connectivity in the corticomuscular system that contributes to functional recovery: in other words, that it can factorize the observed data under the constraint of the connectivity between the observed and source data. Our work thus demonstrates the importance of using not only observed data but also source data to reveal unknown hidden elements.

****

Fundamental limits on the robustness of image classifiers

Zheng Dai,David Gifford

We prove that image classifiers are fundamentally sensitive to small perturbations in their inputs. Specifically, we show that given some image space of n-by-n images, all but a tiny fraction of images in any image class induced over that space can be moved outside that class by adding some perturbation whose p-norm is $O(n^{1/\max{(p,1)}})$, as long as that image class takes up at most half of the image space. We then show that $O(n^{1/\max{(p,1)}})$ is asymptotically optimal. Finally, we show that an increase in the bit depth of the image space leads to a loss in robustness. We supplement our results with a discussion of their implications for vision systems.

****

Stochastic Constrained DRO with a Complexity Independent of Sample Size

Qi Qi,Jiameng Lyu,Kung-sik Chan,Er-Wei Bai,Tianbao Yang

Distributionally Robust Optimization (DRO), as a popular method to train robust models against distribution shifts between training and test sets, has received tremendous attention in recent years. In this paper, we propose and analyze stochastic algorithms that apply to both non-convex and convex losses for solving Kullback–Leibler divergence constrained DRO problem. Compared with existing methods solving this problem, such as primal-dual methods and large mini-batch methods, our stochastic algorithms not only enjoy competitive if not better complexity independent of sample size but also just require a constant batch size at every iteration, which is more practical for broad applications. We establish a nearly optimal complexity bound for finding an ϵ-stationary solution for non-convex losses and an optimal complexity for finding an ϵ-optimal solution for convex losses. Empirical studies demonstrate the effectiveness of the proposed algorithms for solving non-convex and convex constrained DRO problems.

****

Evolve Smoothly, Fit Consistently: Learning Smooth Latent Dynamics For Advection-Dominated Systems

Zhong Yi Wan,Leonardo Zepeda-Nunez,Anudhyan Boral,Fei Sha

We present a data-driven, space-time continuous framework to learn surrogate models for complex physical systems described by advection-dominated partial differential equations. Those systems have slow-decaying Kolmogorov n-width that hinders standard methods, including reduced order modeling, from producing high-fidelity simulations at low cost. In this work, we construct hypernetwork-based latent dynamical models directly on the parameter space of a compact representation network. We leverage the expressive power of the network and a specially designed consistency-inducing regularization to obtain latent trajectories that are both low-dimensional and smooth. These properties render our surrogate models highly efficient at inference time. We show the efficacy of our framework by learning models that generate accurate multi-step rollout predictions at much faster inference speed compared to competitors, for several challenging examples.

****

Dissecting adaptive methods in GANs

Samy Jelassi,David Dobre,Arthur Mensch,Yuanzhi Li,Gauthier Gidel

Adaptive methods are a crucial component widely used for training generative adversarial networks (GANs). While there has been some work to pinpoint the “marginal value of adaptive methods” in standard tasks, it remains unclear why they are still critical for GAN training. In this paper, we formally study how adaptive methods help train GANs; inspired by the grafting method proposed in (Agarwal et al. 2021), we separate the magnitude and direction components of the Adam updates, and graft them to the direction and magnitude of SGDA updates respectively. By considering an update rule with the magnitude of the Adam update and the normalized direction of SGD, we empirically show that the adaptive magnitude of Adam is key for GAN training. This motivates us to have a closer look at the class of normalized stochastic gradient descent ascent (nSGDA) methods in the context of GAN training. We propose a synthetic theoretical framework to compare the performance of nSGDA and SGDA for GAN training with neural networks. We prove that in that setting, GANs trained with nSGDA recover all the modes of the true distribution, whereas the same networks trained with SGDA (and any learning rate configuration) suffer from mode collapse. The critical insight in our analysis is that normalizing the gradients forces the discriminator and generator to be updated at the same pace. We also experimentally show that for several datasets, Adam's performance can be recovered with nSGDA methods.

****

Recycling Scraps: Improving Private Learning by Leveraging Intermediate Checkpoints

Virat Shejwalkar,Arun Ganesh,Rajiv Mathews,Om Thakkar,Abhradeep Guha Thakurta

All state-of-the-art (SOTA) differentially private machine learning (DP ML) methods are iterative in nature, and their privacy analyses allow publicly releasing the intermediate training checkpoints. However, DP ML benchmarks, and even practical deployments, typically use only the final training checkpoint to make predictions. In this work, for the first time, we comprehensively explore various methods that aggregate intermediate checkpoints to improve the utility of DP training. Empirically, we demonstrate that checkpoint aggregations provide significant gains in the prediction accuracy over the existing SOTA for CIFAR10 and StackOverflow datasets, and that these gains get magnified in settings with periodically varying training data distributions. For instance, we improve SOTA StackOverflow accuracies to 22.7\% (+0.43\% absolute) for $\epsilon=8.2$, and 23.84\% (+0.43\%) for $\epsilon=18.9$. Theoretically, we show that uniform tail averaging of checkpoints improves the empirical risk minimization bound compared to the l ast checkpoint of DP-SGD. Lastly, we initiate an exploration into estimating the uncertainty that DP noise adds in the predictions of DP ML models. We prove that, under standard assumptions on the loss function, the sample variance from last few checkpoints provides a good approximation of the variance of the final model of a DP run. Empirically, we show that the last few checkpoints can provide a reasonable lower bound for the variance of a converged DP model.

****

Understanding Influence Functions and Datamodels via Harmonic Analysis
Nikunj Saunshi,Arushi Gupta,Mark Braverman,Sanjeev Arora

Influence functions estimate effect of individual data points on predictions of the model on test data and were adapted to deep learning in \cite{koh2017understanding}. They have been used for detecting data poisoning, detecting helpful and harmful examples, influence of groups of datapoints, etc. Recently, \cite{ilyas2022datamodels} introduced a linear regression method they termed {\em datamodels} to predict the effect of training points on outputs on test data. The current paper seeks to provide a better theoretical understanding of such interesting empirical phenomena. The primary tool is harmonic analysis and the idea of {\em noise stability}. Contributions include: (a) Exact characterization of the learnt datamodel in terms of Fourier coefficients. (b) An efficient method to estimate the residual error and quality of the optimum linear datamodel without having to train the datamodel. (c) New insights into when influences of groups of datapoints may or may not add up linearly.

****

BC-IRL: Learning Generalizable Reward Functions from Demonstrations
Andrew Szot,Amy Zhang,Dhruv Batra,Zsolt Kira,Franziska Meier

How well do reward functions learned with inverse reinforcement learning (IRL) generalize? We illustrate that state-of-the-art IRL algorithms, which maximize a maximum-entropy objective, learn rewards that overfit to the demonstrations. Such rewards struggle to provide meaningful rewards for states not covered by the demonstrations, a major detriment when using the reward to learn policies in new situations. We introduce BC-IRL a new inverse reinforcement learning method that learns reward functions that generalize better when compared to maximum-entropy IRL approaches. In contrast to the MaxEnt framework, which learns to maximize rewards around demonstrations, BC-IRL updates reward parameters such that the policy trained with the new reward matches the expert demonstrations better. We show that BC-IRL learns rewards that generalize better on an illustrative simple task and two continuous robotic control tasks, achieving over twice the success rate of baselines in challenging generalization settings.

****

TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization
Bairu Hou,Jinghan Jia,Yihua Zhang,Guanhua Zhang,Yang Zhang,Sijia Liu,Shiyu Chang

Robustness evaluation against adversarial examples has become increasingly important to unveil the trustworthiness of the prevailing deep models in natural language processing (NLP). However, in contrast to the computer vision domain where the first-order projected gradient descent (PGD) is used as the benchmark approach to generate adversarial examples for robustness evaluation, there lacks a principled first-order gradient-based robustness evaluation framework in NLP. The emerging optimization challenges lie in 1) the discrete nature of textual inputs together with the strong coupling between the perturbation location and the actual content, and 2) the additional constraint that the perturbed text should be fluent and achieve a low perplexity under a language model. These challenges make the development of PGD-like NLP attacks difficult. To bridge the gap, we propose TextGrad, a new attack generator using gradient-driven optimization, supporting high-accuracy and high-quality assessment of adversarial robustness in NLP. Specifically, we address the aforementioned challenges in a unified optimization framework. And we develop an effective convex relaxation method to co-optimize the continuously-relaxed site selection and perturbation variables and leverage an effective sampling method to establish an accurate mapping from the continuous optimization variables to the discrete textual perturbations. Moreover, as a first-order attack generation method, TextGrad can be baked into adversarial traini ng to further improve the robustness of NLP models. Extensive experiments are provided to demonstrate the effectiveness of TextGrad not only in attack generation for robustness evaluation but also in adversarial defense. From the attack perspective, we show that TextGrad achieves remarkable improvements in both the attack success rate and the perplexity score over five state-of-the-art baselines. From the defense perspective, TextGrad-enabled adversarial training yields the most robust NLP model against a wide spectrum of NLP attacks.

****

Robustness for Free: Adversarially Robust Anomaly Detection Through Diffusion Model

Yuanpu Cao,Lu Lin,Jinghui Chen

Deep learning-based anomaly detection models have achieved remarkably high accuracy on commonly used benchmark datasets. However, the robustness of those models may not be satisfactory due to the existence of adversarial examples, which pose significant threats to the practical deployment of deep anomaly detectors. To tackle this issue, we propose an adversarially robust anomaly detector based on the diffusion model. There are two things that make diffusion models a perfect match for our task: 1) the diffusion model itself is a reconstruction-based modeling method whose reconstruction error can serve as a natural indicator of the anomaly score; 2) previous studies have shown that diffusion models can help purify the data for better adversarial robustness. In this work, we highlight that our diffusion model based method gains the adversarial robustness for free: the diffusion model will act both as an anomaly detector and an adversarial defender, thus no extra adversarial training or data purification is needed as in standard robust image classification tasks. We also extend our proposed method for certified robustness to l_2 norm bounded perturbations. Through extensive experiments, we show that our proposed method exhibits outstanding (certified) adversarial robustness while also maintaining equally strong anomaly detection performance on par with the state-of-the-art anomaly detectors on benchmark datasets.

****

ErrorAug: Making Errors to Find Errors in Semantic Segmentation

Devin Guillory,Dequan Wang,Shun Ishizaka,Kazuki Kozuka,Trevor Darrell

In order to develop trustworthy downstream applications for semantic segmentation models, it is important to not only understand the performance of a model on datasets, but to localize areas where the model may produce errors.
Pixel-wise error prediction of semantic segmentation maps is a challenging problem in which prior work relies on complicated image resynthesis pipelines.
We introduce \it{error augmentation}, a framework which enables us to learn robust error detectors by applying data transformations independently on the predicted segmentation maps.
This approach enables direct prediction of pixel-wise error in semantic segmentation maps, an approach explored as a naive baseline in prior works, to achieve state of the art performance.
As a proof-of-concept we propose a series of three simple transformations that generate challenging segmentation errors by swapping pixel predictions within a segmentation map.
Our approach outperforms previous methods of error detection for semantic segmentation across all metrics and improves performance by over 7.8% on AUPR-Error.
Additionally, we show that our approach not only generalizes to unseen test examples, but remains reliable despite significant shifts in the target domain.

****

Kernel Regression with Infinite-Width Neural Networks on Millions of Examples

Ben Adlam,Jaehoon Lee,Shreyas Padhy,Zachary Nado,Jasper Snoek

While kernel regression remains an important practical method, its connection to neural networks as their width becomes large has initiated fresh research. These neural kernels have drastically increased performance on diverse and nonstandard data modalities but require significantly more compute, which previously limited their application to smaller datasets. We address this by massively parallel izing their computation across many GPUs. We combine this with a distributed, preconditioned conjugate gradients algorithm to enable kernel regression at a large scale (i.e. up to 5 million examples). Using this approach, we study scaling laws of several neural kernels across many orders of magnitude for the CIFAR-5m dataset. Using data augmentation to expand the original CIFAR-10 training dataset by a factor of 20, we obtain a test accuracy of 91.2\% (SotA for a pure kernel method). Finally, we explore other data modalities, obtaining results on protein and small molecule prediction tasks that are competitive with SotA methods.

****

Information Plane Analysis for Dropout Neural Networks

Linara Adilova,Bernhard C Geiger,Asja Fischer

The information-theoretic framework promises to explain the predictive power of neural networks. In particular, the information plane analysis, which measures mutual information (MI) between input and representation as well as representation and output, should give rich insights into the training process. This approach, however, was shown to strongly depend on the choice of estimator of the MI. The problem is amplified for deterministic networks if the MI between input and representation is infinite. Thus, the estimated values are defined by the different approaches for estimation, but do not adequately represent the training process from an information-theoretic perspective. In this work, we show that dropout with continuously distributed noise ensures that MI is finite. We demonstrate in a range of experiments that this enables a meaningful information plane analysis for a class of dropout neural networks that is widely used in practice.

****

Fed-Cor: Federated Correlation Test with Secure Aggregation

Lun Wang,Qi Pang,Shuai Wang,Wenting Zheng,Dawn Song

In this paper, we propose the first federated correlation test framework compatible with secure aggregation, namely Fed-Cor. In Fed-Cor, correlation tests are recast as frequency moment estimation problems. To estimate the frequency moments, the clients collaboratively generate a shared projection matrix and then use stable projection to encode the local information in a compact vector. As such encodings can be linearly aggregated, secure aggregation can be applied to conceal the individual updates. We formally establish the security guarantee of Fed-Cor by proving that only the minimum necessary information (i.e., the correlation statistics) is revealed to the server. The evaluation results show that Fed-Cor achieves good accuracy with small client-side computation overhead and performs comparably to the centralized correlation test in several real-world case studies.

****

Feasible Adversarial Robust Reinforcement Learning for Underspecified Environments

John Banister Lanier,Stephen Marcus McAleer,Pierre Baldi,Roy Fox

Robust reinforcement learning (RL) considers the problem of learning policies that perform well in the worst case among a set of possible environment parameter values. In real-world environments, choosing the set of possible values for robust RL can be a difficult task. When that set is specified too narrowly, the agent will be left vulnerable to reasonable parameter values unaccounted for. When specified too broadly, the agent will be too cautious. In this paper, we propose Feasible Adversarial Robust RL (FARR), a novel problem formulation and objective for automatically determining the set of environment parameter values over which to be robust. FARR implicitly defines the set of feasible parameter values as those on which an agent could achieve a benchmark reward given enough training resources. By formulating this problem as a two-player zero-sum game, optimizing the FARR objective jointly produces an adversarial distribution over parameter values with feasible support and a policy robust over this feasible parameter set. We demonstrate that approximate Nash equilibria for this objective can be found using a variation of the PSRO algorithm. Furthermore, we show that an optimal agent trained with FARR is more robust to feasible adversarial parameter selection than with existing minimax, domain-randomization, and regret objectives in a parameterized gridworld and three MuJoCo control environments.

****

Phase2vec: dynamical systems embedding with a physics-informed convolutional network

Matt Ricci,Noa Moriel,Zoe Piran,Mor Nitzan

Dynamical systems are found in innumerable forms across the physical and biological sciences, yet all these systems fall naturally into equivalence classes: conservative or dissipative, stable or unstable, compressible or incompressible. Predicting these classes from data remains an essential open challenge in computational physics on which existing time-series classification methods struggle. Here, we propose, phase2vec, an embedding method that learns high-quality, physically-meaningful representations of low-dimensional dynamical systems without supervision. Our embeddings are produced by a convolutional backbone that extracts geometric features from flow data and minimizes a physically-informed vector field reconstruction loss. The trained architecture can not only predict the equations of unseen data, but also produces embeddings that encode meaningful physical properties of input data (e.g. stability of fixed points, conservation of energy, and the incompressibility of flows) more faithfully than standard blackbox classifiers and state-of-the-art time series classification techniques. We additionally apply our embeddings to the analysis of meteorological data, showing we can detect climatically meaningful features. Collectively, our results demonstrate the viability of embedding approaches for the discovery of dynamical features in physical systems.

****

Learning Harmonic Molecular Representations on Riemannian Manifold

Yiqun Wang,Yuning Shen,Shi Chen,Lihao Wang,Fei YE,Hao Zhou

Molecular representation learning plays a crucial role in AI-assisted drug discovery research. Encoding 3D molecular structures through Euclidean neural networks has become the prevailing method in the geometric deep learning community. However, the equivariance constraints and message passing in Euclidean space may limit the network expressive power. In this work, we propose a Harmonic Molecular Representation learning (HMR) framework, which represents a molecule using the Laplace-Beltrami eigenfunctions of the molecular surface. HMR offers a multi-resolution representation of molecular geometric and chemical properties on 2D Riemannian manifold. We also introduce a harmonic message passing method to realize efficient spectral message passing over the surface manifold for better molecular encoding. Our proposed method shows comparable predictive power to current models in small molecule property prediction, and outperforms the state-of-the-art deep learning models for the rigid protein docking challenge, demonstrating its versatility in molecular representation learning.

****

When is Offline Hyperparameter Selection Feasible for Reinforcement Learning?

Vincent Liu,Prabhat Nagarajan,Andrew Patterson,Martha White

Hyperparameter selection is a critical procedure before deploying reinforcement learning algorithms in real-world applications. However, hyperparameter selection prior to deployment requires selecting policies offline without online execution, which is a significant challenge known as offline policy selection. As yet, there is little understanding about the fundamental limitations of the offline policy selection problem. To contribute to our understanding of this problem, in this paper, we investigate when sample efficient offline policy selection is possible. As off-policy policy evaluation (OPE) is a natural approach for policy selection, the sample complexity of offline policy selection is therefore upper-bounded by the number of samples needed to perform OPE. In addition, we prove that the sample complexity of offline policy selection is also lower-bounded by the sample complexity of OPE. These results imply not only that offline policy selection is effective when OPE is effective, but also that sample efficient policy selection is not possible without additional assumptions that make OPE effective. Moreover, we theoretically study the conditions under which offline policy selection using Fitted Q evaluation (FQE) and the Bellman error is sample efficient. We conclude with an empirical study comparing FQE and Bellman errors for offlin e policy selection.

****

Plansformer: Generating Multi-Domain Symbolic Plans using Transformers

Vishal Pallagani,Bharath Chandra Muppasani,Keerthiram Murugesan,Francesca Rossi,Lior Horesh,Biplav Srivastava,Francesco Fabiano,Andrea Loreggia

Large Language Models (LLMs) have been the subject of active research, significantly advancing the field of Natural Language Processing (NLP). From BERT to BLOOM, LLMs have surpassed state-of-the-art results in various natural language tasks such as question answering, summarization, and text generation. Many ongoing efforts are focused on understanding LLMs' capabilities, including their knowledge of the world, syntax, and semantics. However, extending the textual prowess of LLMs to symbolic reasoning has been slow and predominantly focused on tackling problems related to the mathematical field. In this paper, we explore the use of LLMs for automated planning - a branch of AI concerned with the realization of action sequences (plans) to achieve a goal, typically for execution by intelligent agents, autonomous robots, and unmanned vehicles. We introduce Plansformer; an LLM fine-tuned on planning problems and capable of generating plans with favorable behavior in terms of correctness and length with minimal knowledge-engineering efforts. We also demonstrate the adaptability of Plansformer in solving different planning domains with varying complexities, owing to the transfer learning abilities of LLMs. For one configuration of Plansformer, we achieve ~97\% valid plans, out of which ~95\% are optimal for Towers of Hanoi - a puzzle-solving domain.

****

Greedy Actor-Critic: A New Conditional Cross-Entropy Method for Policy Improvement

Samuel Neumann,Sungsu Lim,Ajin George Joseph,Yangchen Pan,Adam White,Martha White

Many policy gradient methods are variants of Actor-Critic (AC), where a value function (critic) is learned to facilitate updating the parameterized policy (actor). The update to the actor involves a log-likelihood update weighted by the action-values, with the addition of entropy regularization for soft variants. In this work, we explore an alternative update for the actor, based on an extension of the cross entropy method (CEM) to condition on inputs (states). The idea is to start with a broader policy and slowly concentrate around maximal actions, using a maximum likelihood update towards actions in the top percentile per state. The speed of this concentration is controlled by a proposal policy, that concentrates at a slower rate than the actor. We first provide a policy improvement result in an idealized setting, and then prove that our conditional CEM (CCEM) strategy tracks a CEM update per state, even with changing action-values. We empirically show that our Greedy AC algorithm, that uses CCEM for the actor update, performs better than Soft Actor-Critic and is much less sensitive to entropy-regularization.

****

VISION TRANSFORMER FOR MULTIVARIATE TIME- SERIES CLASSIFICATION (VITMTSC)

Prem Shankar Kumar,Ashutosh Joshi,Srinivas Adavi

Multivariate Time-Series Classification (MTSC) is an important issue in many disciplines because of the proliferation of disparate data sources and sensors (economics, retail, health, etc.). Nonetheless, it remains difficult due to the high-dimensionality and richness of data that is regularly updated. We present a Vision Transformer for Multivariate Time-Series Classification (VitMTSC) model that learns latent features from raw time-series data for classification tasks and is applicable to large-scale time-series data with millions of data samples of variable lengths. According to our knowledge, this is the first implementation of the Vision Transformer (ViT) for MTSC. We demonstrate that our approach works on datasets ranging from a few thousand to millions of samples and achieves close to the state-of-the-art (SOTA) results on open datasets. Using click-stream data from a major retail website, we demonstrate that our model can scale to millions of samples and vastly outperform previous neural net-based MTSC models in real-world applications. Our source code is publicly accessible at https://github.co m/mtsc-research/vitmtsc to facilitate further research.

****

Multi-Environment Pretraining Enables Transfer to Action Limited Datasets

David Venuto,Sherry Yang,Pieter Abbeel,Doina Precup,Igor Mordatch,Ofir Nachum

Using massive datasets to train large-scale models has emerged as a dominant approach for broad generalization in natural language and vision applications. In reinforcement learning, however, a key challenge is that available data of sequential decision making is often not annotated with actions - for example, videos of game-play are much more available than sequences of frames paired with the logged game controls. We propose to circumvent this challenge by combining large but sparsely-annotated datasets from a \emph{target} environment of interest with fully-annotated datasets from various other \emph{source} environments. Our method, Action Limited PreTraining (ALPT), leverages the generalization capabilities of inverse dynamics modelling (IDM) to label missing action data in the target environment. We show that utilizing even one additional environment dataset of labelled data during IDM pretraining gives rise to substantial improvements in generating action labels for unannotated sequences. We evaluate our method on benchmark game-playing environments and show that we can significantly improve game performance and generalization capability compared to other approaches, even when using annotated datasets equivalent to only 12 minutes of gameplay.

****

Efficiently Controlling Multiple Risks with Pareto Testing

Bracha Laufer-Goldshtein,Adam Fisch,Regina Barzilay,Tommi S. Jaakkola

Machine learning applications frequently come with multiple diverse objectives and constraints that can change over time. Accordingly, trained models can be tuned with sets of hyper-parameters that affect their predictive behavior (e.g., their run-time efficiency versus error rate). As the number of constraints and hyper-parameter dimensions grow, naively selected settings may lead to sub-optimal and/or unreliable results. We develop an efficient method for calibrating models such that their predictions provably satisfy multiple explicit and simultaneous statistical guarantees (e.g., upper-bounded error rates), while also optimizing any number of additional, unconstrained objectives (e.g., total run-time cost). Building on recent results in distribution-free, finite-sample risk control for general losses, we propose Pareto Testing: a two-stage process which combines multi-objective optimization with multiple hypothesis testing. The optimization stage constructs a set of promising combinations on the Pareto frontier. We then apply statistical testing to this frontier only to identify configurations that have (a) high utility with respect to our objectives, and (b) guaranteed risk levels with respect to our constraints, with specifiably high probability. We demonstrate the effectiveness of our approach to reliably accelerate the execution of large-scale Transformer models in natural language processing (NLP) applications. In particular, we show how Pareto Testing can be used to dynamically configure multiple inter-dependent model attributes—including the number of layers computed before exiting, number of attention heads pruned, or number of text tokens considered—to simultaneously control and optimize various accuracy and cost metrics.

****

Graph Mixup with Soft Alignments

Hongyi Ling,Zhimeng Jiang,Meng Liu,Shuiwang Ji,Na Zou

We study graph data augmentation by mixup, which has been used successfully on images. A key operation of mixup is to compute a convex combination of a pair of inputs. This operation is straightforward for grid-like data, such as images, but challenging for graph data. The key difficulty lies in the fact that different graphs typically have different numbers of nodes, and thus there lacks a node-level correspondence between graphs. In this work, we propose a simple yet effective mixup method for graph classification by soft alignments. Specifically, given a pair of graphs, we explicitly obtain node-level correspondence via computing a soft assignment matrix to match the nodes between two graphs. Based on the so ft assignments, we transform the adjacency and node feature matrices of one graph, so that the transformed graph is aligned with the other graph. In this way, any pair of graphs can be mixed directly to generate an augmented graph. We conduct systematic experiments to show that our method can improve the performance and generalization of graph neural networks (GNNs) on various graph classification tasks. In addition, we show that our method can increase the robustness of GNNs against noisy labels.

****

CNN Compression and Search Using Set Transformations with Width Modifiers on Network Architectures

Bilal Siddiqui,Dakai Zhu

We propose a new approach, based on discrete filter pruning, to adapt off-the-shelf models into an embedded environment. Importantly, we circumvent the usually prohibitive costs of model compression. Our method, Structured Coarse Block Pruning (SCBP), prunes whole CNN kernels using width modifiers applied to a novel transformation of convlayers into superblocks. SCBP uses set representations to construct a rudimentary search to provide candidate networks. To test our approach, the original ResNet architectures serve as the baseline and also provide the 'seeds' for our candidate search. The search produces a configurable number of compressed (derived) models. These derived models are often ~20\% faster and ~50\% smaller than their unmodified counterparts. At the expense of accuracy, the size can become even smaller and the inference latency lowered even further. The unique SCBP transformations yield many new model variants, each with their own trade-offs, and does not require GPU clusters or expert humans for training and design.

****

Learning Interpretable Dynamics from Images of a Freely Rotating 3D Rigid Body

Justice Mason,Christine Allen-Blanchette,Nicholas F Zolman,Elizabeth Davison,Naomi Leonard

In many real-world settings, image observations of freely rotating 3D rigid bodies, such as satellites, may be available when low-dimensional measurements are not. However, the high-dimensionality of image data precludes the use of classical estimation techniques to learn the dynamics and a lack of interpretability reduces the usefulness of standard deep learning methods. In this work, we present a physics-informed neural network model to estimate and predict 3D rotational dynamics from image sequences. We achieve this using a multi-stage prediction pipeline that maps individual images to a latent representation homeomorphic to $\mathbf{SO}(3)$, computes angular velocities from latent pairs, and predicts future latent states using the Hamiltonian equations of motion with a learned representation of the Hamiltonian. We demonstrate the efficacy of our approach on a new rotating rigid-body dataset with sequences of rotating cubes and rectangular prisms with uniform and non-uniform density.

****

NOTELA: A Generalizable Method for Source Free Domain Adaptation

Malik Boudiaf,tom denton,Bart van Merrienboer,Vincent Dumoulin,Eleni Triantafillou

Source-free domain adaptation (SFDA) is a compelling problem as it allows to leverage any off-the-shelf model without requiring access to its original training set and adapts it using only unlabelled data. While several SFDA approaches have recently been proposed, their evaluation focuses on a narrow set of distribution shifts for vision tasks, and their generalizability outside of that scope has not yet been investigated. We put those recent approaches to the test by evaluating them on a new set of challenging---due to extreme covariate and label shift---and naturally-occurring distribution shifts in the audio domain. We study the task of adapting a bird species classifier trained on focalized recordings of bird songs to datasets of passive recordings for various geographical locations. I

nterestingly, we find that some recent SFDA methods underperform doing no adaptation at all. Drawing inspiration from those findings and insights, we propose a new method that improves on noisy student approaches by adjusting the teacher's pseudo-labels through Laplacian regularization. Our approach enjoys increased stability and significantly better performance on several of our proposed distribution shifts. We then look back at SFDA benchmarks in the vision domain and find that our approach is competitive with the state-of-the-art there as well.
****

Characteristic Neural Ordinary Differential Equation
Xingzi Xu,Ali Hasan,Khalil Elkhalil,Jie Ding,Vahid Tarokh

We propose Characteristic-Neural Ordinary Differential Equations (C-NODEs), a framework for extending Neural Ordinary Differential Equations (NODEs) beyond ODEs. While NODE models the evolution of latent variables as the solution to an ODE, C-NODE models the evolution of the latent variables as the solution of a family of first-order partial differential equations (PDEs) along curves on which the PDEs reduce to ODEs, referred to as characteristic curves. This reduction along characteristic curves allows for analyzing PDEs through standard techniques used for ODEs, in particular the adjoint sensitivity method. We also derive C-NODE-based continuous normalizing flows, which describe the density evolution of latent variables along multiple dimensions. Empirical results demonstrate the improvements provided by the proposed method for irregularly sampled time series prediction on MuJoCo, PhysioNet, and Human Activity datasets and classification and density estimation on CIFAR-10, SVHN, and MNIST datasets given a similar computational budget as the existing NODE methods.

The results also provide empirical evidence that the learned curves improve the system efficiency using a lower number of parameters and function evaluations compared with those of the baselines.
****

Fast Sampling of Diffusion Models with Exponential Integrator
Qinsheng Zhang,Yongxin Chen

The past few years have witnessed the great success of Diffusion models~(DMs) in generating high-fidelity samples in generative modeling tasks. A major limitation of the DM is its notoriously slow sampling procedure which normally requires hundreds to thousands of time discretization steps of the learned diffusion process to reach the desired accuracy. Our goal is to develop a fast sampling method for DMs with a much less number of steps while retaining high sample quality. To this end, we systematically analyze the sampling procedure in DMs and identify key factors that affect the sample quality, among which the method of discretization is most crucial. By carefully examining the learned diffusion process, we propose Diffusion Exponential Integrator Sampler~(DEIS). It is based on the Exponential Integrator designed for discretizing ordinary differential equations (ODEs) and leverages a semilinear structure of the learned diffusion process to reduce the discretization error. The proposed method can be applied to any DMs and can generate high-fidelity samples in as few as 10 steps. Moreover, by directly using pre-trained DMs, we achieve state-of-art sampling performance when the number of score function evaluation~(NFE) is limited, e.g., 4.17 FID with 10 NFEs, 2.86 FID with only 20 NFEs on CIFAR10.
****

STay-On-the-Ridge (STON'R): Guaranteed Convergence to Local Minimax Equilibrium in Nonconvex-Nonconcave Games
Constantinos Costis Daskalakis,Noah Golowich,EFSTRATIOS PANTELEIMON SKOULAKIS,Emmanouil Zampetakis

Min-max optimization problems involving nonconvex-nonconcave objectives have found important applications in adversarial training and other multi-agent learning settings. Yet, no known gradient descent-based method is guaranteed to converge to (even local notions of) min-max equilibrium in the nonconvex-nonconcave setting. For all known methods, there exist relatively simple objectives for which they cycle or exhibit other undesirable behavior different from converging to a point, let alone to some game-theoretically meaningful one [Flokas et al. '19, Hsieh et al. '21]. The only known convergence guarantees hold under the strong ass umption that the initialization is very close to a local min-max equilibrium [Wang et al. '19]. Moreover, the afore-described challenges are not just theoretical curiosities. All known methods are unstable in practice, even in simple settings.

We propose the first method that is guaranteed to converge to a local min-max equilibrium for smooth nonconvex-nonconcave objectives. Our method is second-order and provably escapes limit cycles as long as it is initialized at an easy-to-find initial point. Both the definition of our method and its convergence analysis are motivated by the topological nature of the problem. In particular, our method is not designed to decrease some potential function, such as the distance of its iterate from the set of local min-max equilibria or the projected gradient of the objective, but is designed to satisfy a topological property that guarantees the avoidance of cycles and implies its convergence.
****

Federated Representation Learning via Maximal Coding Rate Reduction

Juan Cervino,Navid Naderializadeh,Alejandro Ribeiro

We propose a federated methodology to learn low-dimensional representations from a dataset that is distributed among several clients. In particular, we move away from the commonly-used cross-entropy loss in federated learning, and seek to learn shared low-dimensional representations of the data in a decentralized manner via the principle of maximal coding rate reduction (MCR2). Our proposed method, which we refer to as FLOW, utilizes MCR2 as the objective of choice, hence resulting in representations that are both between-class discriminative and within-class compressible. We theoretically show that our distributed algorithm achieves a first-order stationary point. Moreover, we demonstrate, via numerical experiments, the utility of the learned low-dimensional representations.
****

3D Surface Reconstruction in the Wild by Deforming Shape Priors from Synthetic Data

Nicolai Haeni,Jun-Jee Chao,Volkan Isler

We present a new method for category-specific 3D reconstruction from a single image. A limitation of current color image-based 3D reconstruction models is that they do not generalize across datasets, due to domain shift. In contrast, we show that one can learn to reconstruct objects across datasets by shape priors learned from synthetic 3D data and a point cloud pose canonicalization method. Given a single depth image at test time, we first place this partial point cloud in a canonical pose. Then, we use a neural deformation field in the canonical coordinate frame to reconstruct the 3D surface of the object. Finally, we jointly optimize object pose and 3D shape to fit the partial depth observation. Our approach achieves state-of-the-art reconstruction performance across several real-world datasets, even when trained without ground truth camera poses (which are required by some of the state-of-the-art methods). We further show that our method generalizes to different input modalities, from dense depth images to sparse and noisy LIDAR scans.
****

gDDIM: Generalized denoising diffusion implicit models

Qinsheng Zhang,Molei Tao,Yongxin Chen

Our goal is to extend the denoising diffusion implicit model (DDIM) to general diffusion models~(DMs) besides isotropic diffusions. Instead of constructing a non-Markov noising process as in the original DDIM, we examine the mechanism of DDIM from a numerical perspective. We discover that the DDIM can be obtained by using some specific approximations of the score when solving the corresponding stochastic differential equation. We present an interpretation of the accelerating effects of DDIM that also explains the advantages of a deterministic sampling scheme over the stochastic one for fast sampling. Building on this insight, we extend DDIM to general DMs, coined generalized DDIM (gDDIM), with a small but delicate modification in parameterizing the score network. We validate gDDIM in two non-isotropic DMs: Blurring diffusion model (BDM) and Critically-damped Langevin diffusion model (CLD). We observe more than 20 times acceleration in BDM. In th e CLD, a diffusion model by augmenting the diffusion process with velocity, our algorithm achieves an FID score of 2.26, on CIFAR10, with only 50 number of score function evaluations~(NFEs) and an FID score of 2.86 with only 27 NFEs.

****

Panning for Gold in Federated Learning: Targeted Text Extraction under Arbitrarily Large-Scale Aggregation

Hong-Min Chu,Jonas Geiping,Liam H Fowl,Micah Goldblum,Tom Goldstein

As federated learning (FL) matures, privacy attacks against FL systems in turn become more numerous and complex. Attacks on language models have progressed from recovering single sentences in simple classification tasks to recovering larger parts of user data. Current attacks against federated language models are sequence-agnostic and aim to extract as much data as possible from an FL update - often at the expense of fidelity for any particular sequence. Because of this, current attacks fail to extract any meaningful data under large-scale aggregation. In realistic settings, an attacker cares most about a small portion of user data that contains sensitive personal information, for example sequences containing the phrase "my credit card number is ...". In this work, we propose the first attack on FL that achieves targeted extraction of sequences that contain privacy-critical phrases, whereby we employ maliciously modified parameters to allow the transformer itself to filter relevant sequences from aggregated user data and encode them in the gradient update. Our attack can effectively extract sequences of interest even against extremely large-scale aggregation.

****

Artificial Neuronal Ensembles with Learned Context Dependent Gating

Matthew James Tilley,Michelle Miller,David Freedman

Biological neural networks are capable of recruiting different sets of neurons to encode different memories. However, when training artificial neural networks on a set of tasks, typically, no mechanism is employed for selectively producing anything analogous to these neuronal ensembles. Further, artificial neural networks suffer from catastrophic forgetting, where the network's performance rapidly deteriorates as tasks are learned sequentially. By contrast, sequential learning is possible for a range of biological organisms. We introduce Learned Context Dependent Gating (LXDG), a method to flexibly allocate and recall `artificial neuronal ensembles', using a particular network structure and a new set of regularization terms. Activities in the hidden layers of the network are modulated by gates, which are dynamically produced during training. The gates are outputs of networks themselves, trained with a sigmoid output activation. The regularization terms we have introduced correspond to properties exhibited by biological neuronal ensembles. The first term penalizes low gate sparsity, ensuring that only a specified fraction of the network is used. The second term ensures that previously learned gates are recalled when the network is presented with input from previously learned tasks. Finally, there is a regularization term responsible for ensuring that new tasks are encoded in gates that are as orthogonal as possible from previously used ones. We demonstrate the ability of this method to alleviate catastrophic forgetting on continual learning benchmarks. When the new regularization terms are included in the model along with Elastic Weight Consolidation (EWC) it achieves better performance on the benchmark `permuted MNIST' than with EWC alone. The benchmark `rotated MNIST' demonstrates how similar tasks recruit similar neurons to the artificial neuronal ensemble.

****

Linkless Link Prediction via Relational Distillation

Zhichun Guo,William Shiao,Shichang Zhang,Yozen Liu,Nitesh Chawla,Neil Shah,Tong Zhao

Graph Neural Networks (GNNs) have been widely used on graph data and have shown exceptional performance in the task of link prediction. Despite their effectiveness, GNNs often suffer from high latency due to non-trivial neighborhood data dependency in practical deployments. To address this issue, researchers have proposed methods based on knowledge distillation (KD) to transfer the knowledge from teacher GNNs to student MLPs, which are known to be efficient even with industrial scale data, and have shown promising results on node classification. Nonethel ess, using KD to accelerate link prediction is still unexplored. In this work, we start with exploring two direct analogs of traditional KD for link prediction, i.e., predicted logit-based matching and node representation-based matching. Upon observing direct KD analogs do not perform well for link prediction, we propose a relational KD framework, Linkless Link Prediction (LLP). Unlike simple KD methods that match independent link logits or node representations, LLP distills relational knowledge that is centered around each (anchor) node to the student MLP. Specifically, we propose two matching strategies that complement each other: rank-based matching and distribution-based matching. Extensive experiments demonstrate that LLP boosts the link prediction performance of MLPs with significant margins, and even outperforms the teacher GNNs on 6 out of 9 benchmarks. LLP also achieves a 776.37x speedup in link prediction inference compared to GNNs on the large scale Citation2 dataset.

****

A Differentiable Loss Function for Learning Heuristics in A*

Leah Chrestien,Tomáš Pevný,Antonin Komenda,Stefan Edelkamp

Optimization of heuristic functions for the A* algorithm, realized by deep neural networks, is usually done by minimizing square root loss of estimate of the cost to goal values. This paper argues that this does not necessarily lead to a faster search of A* algorithm since its execution relies on relative values instead of absolute ones. As a mitigation, we propose a L* loss, which upper-bounds the number of excessively expanded states inside the A* search. The L* loss, when used in the optimization of state-of-the-art deep neural networks for automated planning in maze domains like Sokoban and maze with teleports, significantly improves the fraction of solved problems, the quality of founded plans, and reduces the number of expanded states to approximately 50%

****

Understanding Multi-Task Scaling in Machine Translation

Patrick Fernandes,Behrooz Ghorbani,Xavier Garcia,Markus Freitag,Orhan Firat

In this work, we provide a large-scale empirical study of the scaling properties of multilingual (multitask) neural machine translation models. We examine how increases in the model size affect the model performance and investigate the role of the individual task weights on the scaling behavior. We find that these weights only affect the multiplicative factor of the scaling law and in particular, the scaling exponent is unaffected by them. Through a novel joint scaling law formulation, we compute the effective number of parameters allocated to each task and examine the role of language similarity in the scaling behavior of our models. We find minimal evidence that language similarity has any impact. In contrast, ``direction'' of the multilinguality plays a big role, with models translating from multiple languages into English having a larger number of effective parameters per task than their reversed counterparts. Finally, we leverage our observations to predict the performance of multilingual models trained with any language weighting at any scale, greatly reducing efforts required for task balancing in large multitask models. Our findings apply to both in-domain and out-of-domain test sets and to multiple evaluation metrics, such as ChrF and BLEURT.

****

Learning Language Representations with Logical Inductive Bias

Jianshu Chen

Transformer architectures have achieved great success in solving natural language tasks, which learn strong language representations from large-scale unlabeled texts. In this paper, we seek to go further beyond and explore a new logical inductive bias for better language representation learning. Logic reasoning is known as a formal methodology to reach answers from given knowledge and facts. Inspired by such a view, we develop a novel neural architecture named FOLNet (First-Order Logic Network), to encode this new inductive bias. We construct a set of neural logic operators as learnable Horn clauses, which are further forward-chained into a fully differentiable neural architecture (FOLNet). Interestingly, we find that the self-attention module in transformers can be composed by two of our neural logic operators, which probably explains their strong reasoning performance. Our proposed FOLNet has the same input and output interfaces as other pretra

ined models and thus could be pretrained/finetuned by using similar losses. It also allows FOLNet to be used in a plug-and-play manner when replacing other pretrained models. With our logical inductive bias, the same set of ``logic deduction skills'' learned through pretraining are expected to be equally capable of solving diverse downstream tasks. For this reason, FOLNet learns language representations that have much stronger transfer capabilities. Experimental results on several language understanding tasks show that our pretrained FOLNet model outperforms the existing strong transformer-based approaches.
****

AsymQ: Asymmetric Q-loss to mitigate overestimation bias in off-policy reinforcement learning
Qinsheng Zhang,Arjun Krishna,Sehoon Ha,Yongxin Chen
It is well-known that off-policy deep reinforcement learning algorithms suffer from overestimation bias in value function approximation. Existing methods to reduce overestimation bias often utilize multiple value function estimators. Consequently, these methods have a larger time and memory consumption. In this work, we propose a new class of policy evaluation algorithms dubbed, \textbf{AsymQ}, that use asymmetric loss functions to train the Q-value network. Departing from the symmetric loss functions such as mean squared error~(MSE) and Huber loss on the Temporal difference~(TD) error, we adopt asymmetric loss functions of the TD-error to impose a higher penalty on overestimation error. We present one such AsymQ loss called \textbf{Softmax MSE~(SMSE)} that can be implemented with minimal modifications to the standard policy evaluation. Empirically, we show that using SMSE loss helps reduce estimation bias, and subsequently improves policy performance when combined with standard reinforcement learning algorithms. With SMSE, even the Deep Deterministic Policy Gradients~(DDPG) algorithm can achieve performance comparable to that of state-of-the-art methods such as the Twin-Delayed DDPG (TD3) and Soft Actor Critic~(SAC) on challenging environments in the OpenAI Gym MuJoCo benchmark. We additionally demonstrate that the proposed SMSE loss can also boost the performance of Deep Q learning (DQN) in Atari games with discrete action spaces.
****

Movement-to-Action Transformer Networks for Temporal Action Proposal Generation
Xiaodan Hu,Narendra Ahuja
The task of generating temporal action proposals is aimed at identifying temporal intervals containing human actions in untrimmed videos. For arbitrary actions, this requires learning long-range interactions. We propose an end-to-end Movement-and-Action Transformer Network (MatNet) that uses results of human movement studies to encode actions ranging from localized, atomic, body part movements, to longer-range, semantic ones, involving movements of subsets of body parts. In particular, we make direct use of the results of Laban Movement Analysis (LMA). We use LMA-based measures of movements as computational definitions of actions. We input RGB + Flow (I3D) features and 3D pose, compute LMA based low-to-high-level movement features from it, and learn the action proposals by applying two heads on the boundary Transformer and three heads on the proposal Transformer, and using five losses with different weights. We visualize and explain relations between the movement descriptors and attention map of the action proposals. We report results from extensive experiments on the Thumos14, ActivityNet and PKU-MMD datasets, showing that MatNet achieves SOTA or better performance on the temporal action proposal generation task.
****

INSPIRE: A Framework for Integrating Individual User Preferences in Recourse
Prateek Yadav,Peter Hase,Mohit Bansal
Most recourse generation approaches optimize for indirect distance-based metrics like diversity, proximity, and sparsity, or a shared cost function across all users to generate recourse. The latter is an unrealistic assumption because users can have diverse feature preferences which they might be willing to act upon and any changes to any undesirable feature might lead to an impractical recourse. In this work, we propose a novel framework to incorporate the individuality of users in both recourse generation and evaluation procedure by focusing on the cos

t incurred by a user when opting for a recourse. To achieve this, we first propose an objective function, Expected Minimum Cost (EMC) that is based on two key ideas: (1) the user should be comfortable adopting at least one solution when presented with multiple options, and (2) we can approximately optimize for users' satisfaction even when their true cost functions (i.e., costs associated with feature changes) are unknown. EMC samples multiple plausible cost functions based on diverse feature preferences in the population and then finds a recourse set with one good solution for each category of user preferences. We optimize EMC with a novel discrete optimization algorithm, Cost-Optimized Local Search (COLS), that is guaranteed to improve the quality of the recourse set over iterations. Our evaluation framework computes the fraction of satisfied users by simulating each user's cost function and then computing the incurred cost for the provided recourse set. Experimental evaluation on popular real-world datasets demonstrates that our method satisfies up to 25.9% more users compared to strong baselines. Moreover, the human evaluation shows that our recourses are preferred more than twice as often as the strongest baseline.

****

How Does Semi-supervised Learning with Pseudo-labelers Work? A Case Study

Yiwen Kou,Zixiang Chen,Yuan Cao,Quanquan Gu

Semi-supervised learning is a popular machine learning paradigm that utilizes a large amount of unlabeled data as well as a small amount of labeled data to facilitate learning tasks. While semi-supervised learning has achieved great success in training neural networks, its theoretical understanding remains largely open. In this paper, we aim to theoretically understand a semi-supervised learning approach based on pre-training and linear probing. In particular, the semi-supervised learning approach we consider first trains a two-layer neural network based on the unlabeled data with the help of pseudo-labelers. Then it linearly probes the pre-trained network on a small amount of labeled data. We prove that, under a certain toy data generation model and two-layer convolutional neural network, the semisupervised learning approach can achieve nearly zero test loss, while a neural network directly trained by supervised learning on the same amount of labeled data can only achieve constant test loss. Through this case study, we demonstrate a separation between semi-supervised learning and supervised learning in terms of test loss provided the same amount of labeled data.

****

Empowering Graph Representation Learning with Test-Time Graph Transformation

Wei Jin,Tong Zhao,Jiayuan Ding,Yozen Liu,Jiliang Tang,Neil Shah

As powerful tools for representation learning on graphs, graph neural networks (GNNs) have facilitated various applications from drug discovery to recommender systems. Nevertheless, the effectiveness of GNNs is immensely challenged by issues related to data quality, such as distribution shift, abnormal features and adversarial attacks. Recent efforts have been made on tackling these issues from a modeling perspective which requires additional cost of changing model architectures or re-training model parameters. In this work, we provide a data-centric view to tackle these issues and propose a graph transformation framework named GTrans which adapts and refines graph data at test time to achieve better performance. We provide theoretical analysis on the design of the framework and discuss why adapting graph data works better than adapting the model. Extensive experiments have demonstrated the effectiveness of GTrans on three distinct scenarios for eight benchmark datasets where suboptimal data is presented. Remarkably, GTrans performs the best in most cases with improvements up to 2.8%, 8.2% and 3.8% over the best baselines on three experimental settings.

****

Provable Robustness against Wasserstein Distribution Shifts via Input Randomization

Aounon Kumar,Alexander Levine,Tom Goldstein,Soheil Feizi

Certified robustness in machine learning has primarily focused on adversarial perturbations with a fixed attack budget for each sample in the input distribution. In this work, we present provable robustness guarantees on the accuracy of a model under bounded Wasserstein shifts of the data distribution. We show that a s

imple procedure that randomizes the input of the model within a transformation space is provably robust to distributional shifts under that transformation. Our framework allows the datum-specific perturbation size to vary across different points in the input distribution and is general enough to include fixed-sized perturbations as well. Our certificates produce guaranteed lower bounds on the performance of the model for any shift (natural or adversarial) of the input distribution within a Wasserstein ball around the original distribution. We apply our technique to certify robustness against natural (non-adversarial) transformations of images such as color shifts, hue shifts, and changes in brightness and saturation. We obtain strong performance guarantees for the robust model under clearly visible shifts in the input images. Our experiments establish the non-vacuousness of our certificates by showing that the certified lower bound on a robust model's accuracy is higher than the empirical accuracy of an undefended model under a distribution shift. Moreover, our results also imply guaranteed lower bounds (hardness result) on the performance of models trained on so-called "unlearnable" datasets that have been poisoned to interfere with model training. We show that the performance of a robust model is guaranteed to remain above a certain threshold on the test distribution even when the base model is trained on the poisoned dataset.
****
GROOT: Corrective Reward Optimization for Generative Sequential Labeling
Kazuma Hashimoto,Karthik Raman
Sequential labeling is a fundamental NLP task, forming the backbone of many applications.
Supervised learning of Seq2Seq models (like T5) has shown great success on these problems.
However there remains a significant disconnect between the training objectives of these models vs the metrics and desiderata we care about in practical applications.
For example, a practical sequence tagging application may want to optimize for a certain precision-recall trade-off (of the top-k predictions) which is quite different from the standard objective of maximizing the likelihood of the gold labeled sequence.
Thus to bridge this gap, we propose GROOT -- a simple yet effective framework for Generative Reward Optimization Of Text sequences.
GROOT works by training a generative sequential labeling model to match the decoder output distribution with that of the (black-box) reward function.
Using an iterative training regime, we first generate prediction candidates, then correct errors in them, and finally contrast those candidates (based on their reward values).
As demonstrated via extensive experiments on four public benchmarks, GROOT significantly improves all reward metrics.
Furthermore, GROOT also leads to improvements of the overall decoder distribution as evidenced by the quality gains of the top-k candidates.
****
Interpretations of Domain Adaptations via Layer Variational Analysis
Huan-Hsin Tseng,Hsin-Yi Lin,Kuo-Hsuan Hung,Yu Tsao
Transfer learning is known to perform efficiently in many applications empirically, yet limited literature reports the mechanism behind the scene. This study establishes both formal derivations and heuristic analysis to formulate the theory of transfer learning in deep learning. Our framework utilizing layer variational analysis proves that the success of transfer learning can be guaranteed with corresponding data conditions. Moreover, our theoretical calculation yields intuitive interpretations towards the knowledge transfer process. Subsequently, an alternative method for network-based transfer learning is derived. The method shows an increase in efficiency and accuracy for domain adaptation. It is particularly advantageous when new domain data is sufficiently sparse during adaptation. Numerical experiments over diverse tasks validated our theory and verified that our analytic expression achieved better performance in domain adaptation than the gradient descent method.

****

Forget Unlearning: Towards True Data-Deletion in Machine Learning

Rishav Chourasia,Neil Shah,Reza Shokri

Unlearning has emerged as a technique to efficiently erase information of deleted records from learned models. We show, however, that the influence created by the original presence of a data point in the training set can still be detected after running certified unlearning algorithms (which can result in its reconstruction by an adversary). Thus, under realistic assumptions about the dynamics of model releases over time and in the presence of adaptive adversaries, we show that unlearning is not equivalent to data deletion and does not guarantee the "right to be forgotten." We then propose a more robust data-deletion guarantee and show that it is necessary to satisfy differential privacy to ensure true data deletion. Under our notion, we propose an accurate, computationally efficient, and secure data-deletion machine learning algorithm in the online setting based on noisy gradient descent algorithm.

****

Evaluating Unsupervised Denoising Requires Unsupervised Metrics

Adria Marcos Morales,Matan Leibovich,Sreyas Mohan,Joshua Lawrence Vincent,Piyush Haluai,MAI TAN,Peter Crozier,Carlos Fernandez-Granda

Unsupervised denoising is a crucial challenge in real-world imaging applications. Unsupervised deep-learning methods have demonstrated impressive performance on benchmarks based on synthetic noise. However, no metrics are available to evaluate these methods in an unsupervised fashion. This is highly problematic for the many practical applications where ground-truth clean images are not available. In this work, we propose two novel metrics: the unsupervised mean squared error (MSE) and the unsupervised peak signal-to-noise ratio (PSNR), which are computed using only noisy data. We provide a theoretical analysis of these metrics, showing that they are asymptotically consistent estimators of the supervised MSE and PSNR. Controlled numerical experiments with synthetic noise confirm that they provide accurate approximations in practice. We validate our approach on real-world data from two imaging modalities: videos in raw format and transmission electron microscopy. Our results demonstrate that the proposed metrics enable unsupervised evaluation of denoising methods based exclusively on noisy data.

****

Denoising Diffusion Samplers

Francisco Vargas,Will Sussman Grathwohl,Arnaud Doucet

Denoising diffusion models are a popular class of generative models providing state-of-the-art results in many domains. One adds gradually noise to data using a diffusion to transform the data distribution into a Gaussian distribution. Samples from the generative model are then obtained by simulating an approximation of the time-reversal of this diffusion initialized by Gaussian samples. Practically, the intractable score terms appearing in the time-reversed process are approximated using score matching techniques. We explore here a similar idea to sample approximately from unnormalized probability density functions and estimate their normalizing constants. We consider a process where the target density diffuses towards a Gaussian. Denoising Diffusion Samplers (DDS) are obtained by approximating the corresponding time-reversal. While score matching is not applicable in this context, we can leverage many of the ideas introduced in generative modeling for Monte Carlo sampling. Existing theoretical results from denoising diffusion models also provide theoretical guarantees for DDS. We discuss the connections between DDS, optimal control and Schr\"odinger bridges and finally demonstrate DDS experimentally on a variety of challenging sampling tasks.

****

How I Learned to Stop Worrying and Love Retraining

Max Zimmer,Christoph Spiegel,Sebastian Pokutta

Many Neural Network Pruning approaches consist of several iterative training and pruning steps, seemingly losing a significant amount of their performance after pruning and then recovering it in the subsequent retraining phase. Recent works of Renda et al. (2020) and Le & Hua (2021) demonstrate the significance of the learning rate schedule during the retraining phase and propose specific heuristi cs for choosing such a schedule for IMP (Han et al., 2015). We place these findings in the context of the results of Li et al. (2020) regarding the training of models within a fixed training budget and demonstrate that, consequently, the retraining phase can be massively shortened using a simple linear learning rate schedule. Improving on existing retraining approaches, we additionally propose a method to adaptively select the initial value of the linear schedule. Going a step further, we propose similarly imposing a budget on the initial dense training phase and show that the resulting simple and efficient method is capable of outperforming significantly more complex or heavily parameterized state-of-the-art approaches that attempt to sparsify the network during training. These findings not only advance our understanding of the retraining phase, but more broadly question the belief that one should aim to avoid the need for retraining and reduce the negative effects of ‘hard’ pruning by incorporating the sparsification process into the standard training.

****

The Value of Out-of-distribution Data

Ashwin De Silva,Rahul Ramesh,Carey Priebe,Pratik Chaudhari,Joshua T Vogelstein

More data is expected to help us generalize to a task. But real datasets can contain out-of-distribution (OOD) data; this can come in the form of heterogeneity such as intra-class variability but also in the form of temporal shifts or concept drifts. We demonstrate a counter-intuitive phenomenon for such problems: generalization error of the task can be a non-monotonic function of the number of OOD samples; a small number of OOD samples can improve generalization but if the number of OOD samples is beyond a threshold, then the generalization error can deteriorate. We also show that if we know which samples are OOD, then using a weighted objective between the target and OOD samples ensures that the generalization error decreases monotonically. We demonstrate and analyze this phenomenon using linear classifiers on synthetic datasets and medium-sized neural networks on vision benchmarks such as MNIST, CIFAR-10, CINIC-10, PACS, and DomainNet, and observe the effect data augmentation, hyperparameter optimization, and pre-training have on this behavior.

****

Factors Influencing Generalization in Chaotic Dynamical Systems

Luã Streit,Vikram Voleti,Tegan Maharaj

Many real-world systems exhibit chaotic behaviour, for example: weather, fluid dynamics, stock markets, natural ecosystems, and disease transmission. While chaotic systems are often thought to be completely unpredictable, in fact there are patterns within and across that experts frequently describe and contrast qualitatively. We hypothesize that given the right supervision / task definition, representation learning systems will be able to pick up on these patterns, and successfully generalize both in- and out-of-distribution (OOD).

Thus, this work explores and identifies key factors which lead to good generalization. We observe a variety of interesting phenomena, including: learned representations transfer much better when fine-tuned vs. frozen; forecasting appears to be the best pre-training task; OOD robustness falls off very quickly outside the training distribution; recurrent architectures generally outperform others on OOD generalization.

Our findings are of interest to any domain of prediction where chaotic dynamics play a role.

****

Interpretable Geometric Deep Learning via Learnable Randomness Injection

Siqi Miao,Yunan Luo,Mia Liu,Pan Li

Point cloud data is ubiquitous in scientific fields. Recently, geometric deep learning (GDL) has been widely applied to solve prediction tasks with such data. However, GDL models are often complicated and hardly interpretable, which poses concerns to scientists who are to deploy these models in scientific analysis and experiments. This work proposes a general mechanism, learnable randomness injection (LRI), which allows building inherently interpretable models based on general GDL backbones. LRI-induced models, once trained, can detect the points in the point cloud data that carry information indicative of the prediction label. We a lso propose four datasets from real scientific applications that cover the domains of high-energy physics and biochemistry to evaluate the LRI mechanism. Compared with previous post-hoc interpretation methods, the points detected by LRI align much better and stabler with the ground-truth patterns that have actual scientific meanings. LRI is grounded by the information bottleneck principle, and thus LRI-induced models are also more robust to distribution shifts between training and test scenarios. Our code and datasets are available at https://github.com/Graph-COM/LRI.

****

Koopman Operator Learning for Accelerating Quantum Optimization and Machine Learning

Di Luo,Jiayu Shen,Rumen Dangovski,Marin Soljacic

Finding efficient optimization methods plays an important role for quantum optimization and quantum machine learning on near-term quantum computers. While backpropagation on classical computers is computationally efficient, obtaining gradients on quantum computers is not, because the computational complexity scales linearly with the number of parameters and measurements. In this paper, we connect Koopman operator theory, which has been successful in predicting nonlinear dynamics, with natural gradient methods in quantum optimization. We propose a data-driven approach using Koopman operator learning to accelerate quantum optimization and quantum machine learning. We develop two new families of methods: the sliding window dynamic mode decomposition (DMD) and the neural DMD for efficiently updating parameters on quantum computers. We show that our methods can predict gradient dynamics on quantum computers and accelerate the quantum variational eigensolver used in quantum optimization, as well as quantum machine learning. We further implement the learning algorithms on a real quantum computer and demonstrate their practical effectiveness.

****

GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure

Tennison Liu,Zhaozhi Qian,Jeroen Berrevoets,Mihaela van der Schaar

Deep generative models learn highly complex and non-linear representations to generate realistic synthetic data. While they have achieved notable success in computer vision and natural language processing, similar advances have been less demonstrable in the tabular domain. This is partially because generative modelling of tabular data entails a particular set of challenges, including heterogeneous relationships, limited number of samples, and difficulties in incorporating prior knowledge. Additionally, unlike their counterparts in image and sequence domain, deep generative models for tabular data almost exclusively employ fully-connected layers, which encode weak inductive biases about relationships between inputs. Real-world data generating processes can often be represented using relational structures, which encode sparse, heterogeneous relationships between variables. In this work, we learn and exploit relational structure underlying tabular data to better model variable dependence, and as a natural means to introduce regularization on relationships and include prior knowledge. Specifically, we introduce GOGGLE, an end-to-end message passing scheme that jointly learns the relational structure and corresponding functional relationships as the basis of generating synthetic samples. Using real-world datasets, we provide empirical evidence that the proposed method is effective in generating realistic synthetic data and exploiting domain knowledge for downstream tasks.

****

A Reproducible and Realistic Evaluation of Partial Domain Adaptation Methods

Tiago Salvador,Kilian FATRAS,Ioannis Mitliagkas,Adam M Oberman

Unsupervised Domain Adaptation (UDA) aims at classifying unlabeled target images leveraging source labeled ones. In this work, we consider the Partial Domain Adaptation (PDA) variant, where we have extra source classes not present in the target domain. Most successful algorithms use model selection strategies that rely on target labels to find the best hyper-parameters and/or models along training. However, these strategies violate the main assumption in PDA: only unlabeled target domain samples are available. Moreover, there are also inconsistencies in the experimental settings - architecture, hyper-parameter tuning, number of runs

- yielding unfair comparisons. The main goal of this work is to provide a realistic evaluation of PDA methods with the different model selection strategies under a consistent evaluation protocol. We evaluate 7 representative PDA algorithms on 2 different real-world datasets using 7 different model selection strategies. Our two main findings are: (i) without target labels for model selection, the accuracy of the methods decreases up to 30 percentage points; (ii) only one method and model selection pair performs well on both datasets. Experiments were performed with our PyTorch framework, BenchmarkPDA, which we open source.

****

Progressive Prompts: Continual Learning for Language Models

Anastasia Razdaibiedina,Yuning Mao,Rui Hou,Madian Khabsa,Mike Lewis,Amjad Almahairi

We introduce Progressive Prompts – a simple and efficient approach for continual learning in language models. Our method allows forward transfer and resists catastrophic forgetting, without relying on data replay or a large number of task-specific parameters. Progressive Prompts learns a new soft prompt for each task and sequentially concatenates it with the previously learned prompts, while keeping the base model frozen. Experiments on standard continual learning benchmarks show that our approach outperforms state-of-the-art methods, with an improvement >20% in average test accuracy over the previous best-preforming method on T5 model. We also explore a more challenging continual learning setup with longer sequences of tasks and show that Progressive Prompts significantly outperforms prior methods.

****

Differentiable Rendering with Reparameterized Volume Sampling

Kirill Struminsky,Oleg Desheulin

We propose an alternative rendering algorithm for neural radiance fields based on importance sampling. In view synthesis, a neural radiance field approximates underlying density and radiance fields based on a sparse set of views of a scene. To generate a pixel of a novel view, it marches a ray through the pixel and computes a weighted sum of radiance emitted from a dense set of ray points. This rendering algorithm is fully differentiable and facilitates gradient-based optimization of the fields. However, in practice, only a tiny opaque portion of the ray contributes most of the radiance to the sum. Therefore, we can avoid computing radiance in the rest part. In this work, we use importance sampling to pick non-transparent points on the ray. Specifically, we generate samples according to the probability distribution induced by the density field. Our main contribution is the reparameterization of the sampling algorithm. It allows end-to-end learning with gradient descent as in the original rendering algorithm. With our approach, we can optimize a neural radiance field with just a few radiance field evaluations per ray. As a result, we alleviate the costs associated with the color component of the neural radiance field.

****

Deep Learning From Crowdsourced Labels: Coupled Cross-Entropy Minimization, Identifiability, and Regularization

Shahana Ibrahim,Tri Nguyen,Xiao Fu

Using noisy crowdsourced labels from multiple annotators, a deep learning-based end-to-end (E2E) system aims to learn the label correction mechanism and the neural classifier simultaneously. To this end, many E2E systems concatenate the neural classifier with multiple annotator-specific label confusion layers and co-train the two parts in a parameter-coupled manner. The formulated coupled cross-entropy minimization (CCEM)-type criteria are intuitive and work well in practice. Nonetheless, theoretical understanding of the CCEM criterion has been limited. The contribution of this work is twofold: First, performance guarantees of the CCEM criterion are presented. Our analysis reveals for the first time that the CCEM can indeed correctly identify the annotators' confusion characteristics and the desired ``ground-truth'' neural classifier under realistic conditions, e.g., when only incomplete annotator labeling and finite samples are available. Second, based on the insights learned from our analysis, two regularized variants of the CCEM are proposed. The regularization terms provably enhance the identifiabi lity of the target model parameters in various more challenging cases. A series of synthetic and real data experiments are presented to showcase the effectiveness of our approach.

****

Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference

Pierre Glaser,Michael Arbel,Arnaud Doucet,Arthur Gretton

We introduce two Synthetic Likelihood methods for Simulation-Based Inference (SBI), to conduct either amortized or targeted inference from experimental observations when a high-fidelity simulator is available. Both methods learn a Conditional Energy-Based Model (EBM) of the likelihood using synthetic data generated by the simulator, conditioned on parameters drawn from a proposal distribution. The learned likelihood can then be combined with any prior to obtain a posterior estimate, from which samples can be drawn using MCMC.

Our methods uniquely combine a flexible Energy-Based Model and the minimization of a KL loss: this is in contrast to other synthetic likelihood methods, which either rely on normalizing flows, or minimize score-based objectives; choices that come with known pitfalls. Our first method, Amortized Unnormalized Neural Likelihood Estimation (AUNLE), introduces a tilting trick during training that allows to perform inference using efficient MCMC techniques. Our second method, Sequential UNLE (SUNLE), employs a doubly intractable approach in order to re-use simulation data and improve posterior accuracy for a specific observation.

We demonstrate the properties of both methods on a range of synthetic datasets, and apply it to a neuroscience model of the pyloric network in the crab, matching the performance of other synthetic likelihood methods at a fraction of the simulation budget.

****

Just Avoid Robust Inaccuracy: Boosting Robustness Without Sacrificing Accuracy

Yannick Merkli,Pavol Bielik,PETAR TSANKOV,Martin Vechev

While current methods for training robust deep learning models optimize robust accuracy, they significantly reduce natural accuracy, hindering their adoption in practice. Further, the resulting models are often both robust and inaccurate on numerous samples, providing a false sense of safety for those. In this work, we extend prior works in three main directions. First, we explicitly train the models to jointly maximize robust accuracy and minimize robust inaccuracy. Second, since the resulting models are trained to be robust only if they are accurate, we leverage robustness as a principled abstain mechanism. Finally, this abstain mechanism allows us to combine models in a compositional architecture that significantly boosts overall robustness without sacrificing accuracy. We demonstrate the effectiveness of our approach for empirical and certified robustness on six recent state-of-the-art models and four datasets. For example, on CIFAR-10 with $\epsilon_\infty = 1/255$, we successfully enhanced the robust accuracy of a pre-trained model from 26.2% to 87.8% while even slightly increasing its natural accuracy from 97.8% to 98.0%.

****

Projective Proximal Gradient Descent for Nonconvex Nonsmooth Optimization: Fast Convergence Without Kurdyka-Lojasiewicz (KL) Property

Yingzhen Yang,Ping Li

Nonconvex and nonsmooth optimization problems are important and challenging for statistics and machine learning. In this paper, we propose Projected Proximal Gradient Descent (PPGD) which solves a class of nonconvex and nonsmooth optimization problems, where the nonconvexity and nonsmoothness come from a nonsmooth regularization term which is nonconvex but piecewise convex. In contrast with existing convergence analysis of accelerated PGD methods for nonconvex and nonsmooth problems based on the Kurdyka-\L{}ojasiewicz (K\L{}) property, we provide a new theoretical analysis showing local fast convergence of PPGD. It is proved that PPGD achieves a fast convergence rate of $O(1/k^2)$ when the iteration number $k \ge k_0$ for a finite k_0 on a class of nonconvex and nonsmooth problems under mild assumptions, which is locally the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continu ous gradient. Experimental results demonstrate the effectiveness of PPGD.

****

First Steps Toward Understanding the Extrapolation of Nonlinear Models to Unseen Domains

Kefan Dong,Tengyu Ma

Real-world machine learning applications often involve deploying neural networks to domains that are not seen in the training time. Hence, we need to understand the extrapolation of \textit{nonlinear} models---under what conditions on the distributions and function class, models can be guaranteed to extrapolate to new test distributions. The question is very challenging because even two-layer neural networks cannot be guaranteed to extrapolate outside the support of the training distribution without further assumptions on the domain shift. This paper makes some initial steps towards analyzing the extrapolation of nonlinear models for structured domain shift. We primarily consider settings where the \textit{marginal} distribution of each coordinate of the data (or subset of coordinates) do not shift significantly across the training and test distributions, but the joint distribution may have a much bigger shift. We prove that the family of nonlinear models of the form $f(x)=\sum f_i(x_i)$, where f_i is an \emph{arbitrary} function on the subset of features x_i, can extrapolate to unseen distributions, if the covariance of the features is well-conditioned. To the best of our knowledge, this is the first result that goes beyond linear models and the bounded density ratio assumption, even though the assumptions on the distribution shift and function class are stylized.

****

A Kernel-Based View of Language Model Fine-Tuning

Sadhika Malladi,Alexander Wettig,Dingli Yu,Danqi Chen,Sanjeev Arora

It has become standard to solve NLP tasks by fine-tuning pre-trained language models (LMs), especially in low-data settings. There is minimal theoretical understanding of empirical success, e.g., why fine-tuning a model with 10^8 or more parameters on a couple dozen training points does not result in overfitting. We investigate whether the Neural Tangent Kernel (NTK)--which originated as a model to study the gradient descent dynamics of infinitely wide networks with suitable random initialization--describes fine-tuning of pre-trained LMs. This study was inspired by the decent performance of NTK for computer vision tasks (Wei et al., 2022). We also extend the NTK formalism to fine-tuning with Adam. We present extensive experiments that suggest that once the task is formulated as a masked language modeling problem through prompting, the NTK lens can often reasonably describe the model updates during fine-tuning with both SGD and Adam.
This kernel view also suggests an explanation for success of parameter-efficient subspace-based fine-tuning methods. Finally, we suggest a path toward a formal explanation for our findings via Tensor Programs (Yang, 2020).

****

Compositionality with Variation Reliably Emerges in Neural Networks

Henry Conklin,Kenny Smith

Human languages enable robust generalization, letting us leverage our prior experience to communicate about novel meanings. This is partly due to language being compositional, where the meaning of a whole expression is a function of its parts. Natural languages also exhibit extensive variation, encoding meaning predictably enough to enable generalization without limiting speakers to one and only one way of expressing something. Previous work looking at the languages that emerge between neural networks in a communicative task has shown languages that enable robust communication and generalization reliably emerge. Despite this those languages score poorly on existing measures of compositionality leading to claims that a language's degree of compositionality has little bearing on how well it can generalise. We argue that the languages that emerge between networks are in fact straightforwardly compositional, but with a degree of natural language-like variation that can obscure their compositionality from existing measures. We introduce 4 measures of linguistic variation and show that early in training measures of variation correlate with generalization performance, but that this effect goes away over time as the languages that emerge become regular enough to gener alize robustly. Like natural languages, emergent languages appear able to support a high degree of variation while retaining the generalizability we expect from compositionality. In an effort to decrease the variability of emergent languages we show how reducing a model's capacity results in greater regularity, in line with claims about factors shaping the emergence of regularity in human language.

****

Systematic Rectification of Language Models via Dead-end Analysis

Meng Cao,Mehdi Fatemi,Jackie CK Cheung,Samira Shabanian

With adversarial or otherwise normal prompts, existing large language models (LLM) can be pushed to generate toxic discourses. One way to reduce the risk of LLMs generating undesired discourses is to alter the training of the LLM. This can be very restrictive due to demanding computation requirements. Other methods rely on rule-based or prompt-based token elimination, which are limited as they dismiss future tokens and the overall meaning of the complete discourse. Here, we center detoxification on the probability that the finished discourse is ultimately considered toxic. That is, at each point, we advise against token selections proportional to how likely a finished text from this point will be toxic. To this end, we formally extend the dead-end theory from the recent reinforcement learning (RL) literature to also cover uncertain outcomes. Our approach, called rectification, utilizes a separate but significantly smaller model for detoxification, which can be applied to diverse LLMs as long as they share the same vocabulary. Importantly, our method does not require access to the internal representations of the LLM, but only the token probability distribution at each decoding step. We believe this is important since many LLMs today are hosted in servers and only accessible through APIs. When applied to various LLMs, including GPT-3, our approach generates notably better results compared to the base LLMs and other techniques in terms of the overall language and detoxification performance.

****

Model-free Reinforcement Learning that Transfers Using Random Reward Features

Boyuan Chen,Chuning Zhu,Pulkit Agrawal,Kaiqing Zhang,Abhishek Gupta

Favorable reinforcement learning (RL) algorithms should not only be able to synthesize controller for complex tasks, but also transfer across various such tasks. Classical model-free RL algorithms like Q-learning can be made stable, and has the potential to solve complicated tasks individually. However, rewards are key supervision signals in model-free approaches, making it challenging in general to transfer across multiple tasks with different reward functions. On the other hand, model-based RL algorithms, naturally transfers to various reward functions if the transition dynamics are learned well. Unfortunately, model-learning usually suffers from high dimensional observations and/or long horizons due to the challenges of compounding error. In this work, we propose a new way to transfer behaviors across problems with different reward functions that enjoy the best of both worlds. Specifically, we develop a model-free approach that implicitly learns the model without constructing the transition dynamics. This is achieved by using random features to generate reward functions in training, and incorporating model predictive control with open-loop policies in online planning. We show that the approach enables fast adaptation to problems with completely new reward functions, while scaling to high dimensional observations and long horizons. Moreover, our method can easily be trained on large offline datasets, and be quickly deployed on new tasks with good performance, making it more widely applicable than typical model-free and model-based RL methods. We evaluate the superior performance of our algorithm in a variety of RL and robotics domains.

****

Multiple sequence alignment as a sequence-to-sequence learning problem

Edo Dotan,Yonatan Belinkov,Oren Avram,Elya Wygoda,Noa Ecker,Michael Alburquerque,Omri Keren,Gil Loewenthal,Tal Pupko

The sequence alignment problem is one of the most fundamental problems in bioinformatics and a plethora of methods were devised to tackle it. Here we introduce BetaAlign, a methodology for aligning sequences using an NLP approach. BetaAlign accounts for the possible variability of the evolutionary process among differe nt datasets by using an ensemble of transformers, each trained on millions of samples generated from a different evolutionary model. Our approach leads to alignment accuracy that is similar and often better than commonly used methods, such as MAFFT, DIALIGN, ClustalW, T-Coffee, PRANK, and MUSCLE.

****

Fair Graph Message Passing with Transparency
Zhimeng Jiang,Xiaotian Han,Chao Fan,Zirui Liu,Na Zou,Ali Mostafavi,Xia Hu

Recent advanced works achieve fair representations and predictions through regularization, adversarial debiasing, and contrastive learning in graph neural networks (GNNs). These methods \textit{implicitly} encode the sensitive attribute information in the well-trained model weight via \textit{backward propagation}. In practice, we not only pursue a fair machine learning model but also lend such fairness perception to the public. For current fairness methods,
how the sensitive attribute information usage makes the model achieve fair prediction still remains a black box. In this work, we first propose the concept \textit{transparency} to describe \textit{whether} the model embraces the ability of lending fairness perception to the public \textit{or not}. Motivated by the fact that current fairness models lack of transparency, we aim to pursue a fair machine learning model with transparency via \textit{explicitly} rendering sensitive attribute usage for fair prediction in \textit{forward propagation} . Specifically, we develop an effective and transparent \textsf{F}air \textsf{M}essage \textsf{P}assing (FMP) scheme adopting sensitive attribute information in forward propagation. In this way, FMP explicitly uncovers how sensitive attributes influence final prediction. Additionally, FMP scheme can aggregate useful information from neighbors and mitigate bias in a unified framework to simultaneously achieve graph smoothness and fairness objectives. An acceleration approach is also adopted to improve the efficiency of FMP. Experiments on node classification tasks demonstrate that the proposed FMP outperforms the state-of-the-art baselines in terms of fairness and accuracy on three real-world datasets. The code is available in {\color{blue}\url{https://anonymous.4open.science/r/FMP-AD84}}.

****

FedExP: Speeding Up Federated Averaging via Extrapolation
Divyansh Jhunjhunwala,Shiqiang Wang,Gauri Joshi

Federated Averaging (FedAvg) remains the most popular algorithm for Federated Learning (FL) optimization due to its simple implementation, stateless nature, and privacy guarantees combined with secure aggregation. Recent work has sought to generalize the vanilla averaging in FedAvg to a generalized gradient descent step by treating client updates as pseudo-gradients and using a server step size. While the use of a server step size has been shown to provide performance improvement theoretically, the practical benefit of the server step size has not been seen in most existing works. In this work, we present FedExP, a method to adaptively determine the server step size in FL based on dynamically varying pseudo-gradients throughout the FL process. We begin by considering the overparameterized convex regime, where we reveal an interesting similarity between FedAvg and the Projection Onto Convex Sets (POCS) algorithm. We then show how FedExP can be motivated as a novel extension to the extrapolation mechanism that is used to speed up POCS. Our theoretical analysis later also discusses the implications of FedExP in underparameterized and non-convex settings. Experimental results show that FedExP consistently converges faster than FedAvg and competing baselines on a range of realistic FL datasets.

****

Graph Neural Networks Are More Powerful Than we Think
Charilaos Kanatsoulis,Alejandro Ribeiro

Graph Neural Networks (GNNs) are powerful convolutional architectures that have shown remarkable performance in various node-level and graph-level tasks. Despite their success, the common belief is that the expressive power of standard GNNs is limited and that they are at most as discriminative as the Weisfeiler-Lehman (WL) algorithm. In this paper we argue the opposite and show that the WL algorithm is the upper bound only when the input to the GNN is the vector of all ones. In this direction, we derive an alternative analysis that employs linear algebr aic tools and characterize the representational power of GNNs with respect to the eigenvalue decomposition of the graph operators. We show that GNNs can distinguish between any graphs that differ in at least one eigenvalue and design simple GNN architectures that are provably more expressive than the WL algorithm. Thorough experimental analysis on graph isomorphism and graph classification datasets corroborates our theoretical results and demonstrates the effectiveness of the proposed architectures.

****

A Mixture-of-Expert Approach to RL-based Dialogue Management

Yinlam Chow,Azamat Tulepbergenov,Ofir Nachum,Dhawal Gupta,Moonkyung Ryu,Mohammad Ghavamzadeh,Craig Boutilier

Despite recent advancements in language models (LMs), their application to dialogue management (DM) problems and ability to carry on rich conversations remain a challenge. We use reinforcement learning (RL) to develop a dialogue agent that avoids being short-sighted (outputting generic utterances) and maximizes overall user satisfaction. Most existing RL approaches to DM train the agent at the word-level, and thus, have to deal with a combinatorially complex action space even for a medium-size vocabulary. As a result, they struggle to produce a successful and engaging dialogue even if they are warm-started with a pre-trained LM. To address this issue, we develop a RL-based DM using a novel mixture of expert language model (MoE-LM) that consists of (i) a LM capable of learning diverse semantics for conversation histories, (ii) a number of specialized LMs (or experts) capable of generating utterances corresponding to a particular attribute or personality, and (iii) a RL-based DM that performs dialogue planning with the utterances generated by the experts. Our MoE approach provides greater flexibility to generate sensible utterances with different intents and allows RL to focus on conversational-level DM. We compare it with SOTA baselines on open-domain dialogues and demonstrate its effectiveness both in terms of the diversity and sensibility of the generated utterances and the overall DM performance.

****

f-DM: A Multi-stage Diffusion Model via Progressive Signal Transformation

Jiatao Gu,Shuangfei Zhai,Yizhe Zhang,Miguel Ángel Bautista,Joshua M. Susskind

Diffusion models (DMs) have recently emerged as SoTA tools for generative modeling in various domains. Standard DMs can be viewed as an instantiation of hierarchical variational autoencoders (VAEs) where the latent variables are inferred from input-centered Gaussian distributions with fixed scales and variances. Unlike VAEs, this formulation constrains DMs from changing the latent spaces and learning abstract representations. In this work, we propose f-DM, a generalized family of DMs which allows progressive signal transformation. More precisely, we extend DMs to incorporate a set of (hand-designed or learned) transformations, where the transformed input is the mean of each diffusion step. We propose a generalized formulation and derive the corresponding de-noising objective with a modified sampling algorithm. As a demonstration, we apply f-DM in image generation tasks with a range of functions, including down-sampling, blurring, and learned transformations based on the encoder of pretrained VAEs. In addition, we identify the importance of adjusting the noise levels whenever the signal is sub-sampled and propose a simple rescaling recipe. f-DM can produce high-quality samples on standard image generation benchmarks like FFHQ, AFHQ, LSUN, and ImageNet with better efficiency and semantic interpretation.

****

An Empirical Study of the Neural Contextual Bandit Algorithms

Bowen Wei,Yiling Jia,Hongning Wang

Recent advances in representation learning have made significant influences on solutions of contextual bandit problems. Neural bandit algorithms have been actively developed and reported to gain extraordinary performance improvement against classical bandit algorithms in numerous papers. However, there lacks a comprehensive comparison among the existing neural bandit algorithms, and it is still not clear whether or when they can succeed in complex real-world problems. In this work, we present an inclusive empirical study on three different categories of existing neural bandit algorithms on several real-world datasets. The results sh ow that such algorithms are highly competitive against their classical counterparts in most cases, however the advantage is not consistent. The results also reveal crucial challenges for future research in neural bandit algorithms.

****

Backpropagation at the Infinitesimal Inference Limit of Energy-Based Models: Unifying Predictive Coding, Equilibrium Propagation, and Contrastive Hebbian Learning

Beren Millidge,Yuhang Song,Tommaso Salvatori,Thomas Lukasiewicz,Rafal Bogacz

How the brain performs credit assignment is a fundamental unsolved problem in neuroscience. Many `biologically plausible' algorithms have been proposed, which compute gradients that approximate those computed by backpropagation (BP), and which operate in ways that more closely satisfy the constraints imposed by neural circuitry. Many such algorithms utilize the framework of energy-based models (EBMs), in which all free variables in the model are optimized to minimize a global energy function. However, in the literature, these algorithms exist in isolation and no unified theory exists linking them together. Here, we provide a comprehensive theory of the conditions under which EBMs can approximate BP, which lets us unify many of the BP approximation results in the literature (namely, predictive coding, equilibrium propagation, and contrastive Hebbian learning) and demonstrate that their approximation to BP arises from a simple and general mathematical property of EBMs at free-phase equilibrium. This property can then be exploited in different ways with different energy functions, and these specific choices yield a family of BP-approximating algorithms, which both includes the known results in the literature and can be used to derive new ones.

****

A Theoretical Framework for Inference and Learning in Predictive Coding Networks

Beren Millidge,Yuhang Song,Tommaso Salvatori,Thomas Lukasiewicz,Rafal Bogacz

Predictive coding (PC) is an influential theory in computational neuroscience, which argues that the cortex forms unsupervised world models by implementing a hierarchical process of prediction error minimization. PC networks (PCNs) are trained in two phases. First, neural activities are updated to optimize the network's response to external stimuli. Second, synaptic weights are updated to consolidate this change in activity --- an algorithm called \emph{prospective configuration}. While previous work has shown how in various limits, PCNs can be found to approximate backpropagation (BP), recent work has demonstrated that PCNs operating in this standard regime, which does not approximate BP, nevertheless obtain competitive training and generalization performance to BP-trained networks while outperforming them on various tasks. However, little is understood theoretically about the properties and dynamics of PCNs in this regime. In this paper, we provide a comprehensive theoretical analysis of the properties of PCNs trained with prospective configuration. We first derive analytical results concerning the inference equilibrium for PCNs and a previously unknown close connection relationship to target propagation (TP). Secondly, we provide a theoretical analysis of learning in PCNs as a variant of generalized expectation-maximization and use that to prove the convergence of PCNs to critical points of the BP loss function, thus showing that deep PCNs can, in theory, achieve the same generalization performance as BP, while maintaining their unique advantages.

****

Causally-guided Regularization of Graph Attention improves Generalizability

Alexander P Wu,Thomas Markovich,Bonnie Berger,Nils Yannick Hammerla,Rohit Singh

Graph attention networks estimate the relational importance of node neighbors to aggregate relevant information over local neighborhoods for a prediction task. However, the inferred attentions are vulnerable to spurious correlations and connectivity in the training data, hampering the generalizability of the model. We introduce CAR, a general-purpose regularization framework for graph attention networks. Embodying a causal inference approach, CAR aligns the attention mechanism with the causal effects of active interventions on graph connectivity in a scalable manner. CAR is compatible with a variety of graph attention architectures, and we show that it systematically improves generalizability on various node classification tasks. Our ablation studies indicate that CAR hones in on the aspec ts of graph structure most pertinent to the prediction (e.g., homophily), and does so more effectively than alternative approaches. Finally, we also show that CAR enhances interpretability of attention weights by accentuating node-neighbor relations that point to causal hypotheses. For social media network-sized graphs, a CAR-guided graph rewiring approach could allow us to combine the scalability of graph convolutional methods with the higher performance of graph attention.

****

On a Benefit of Masked Language Model Pretraining: Robustness to Simplicity Bias
Ting-Rui Chiang

Despite the success of pretrained masked language models (MLM), why MLM pretraining is useful is still a question not fully answered. In this work we theoretically and empirically show that MLM pretraining makes models robust to lexicon-level spurious features, partly answering the question. Our explanation is that MLM pretraining may alleviate problems brought by simplicity bias (Shahet al., 2020), which refers to the phenomenon that a deep model tends to rely excessively on simple features. In NLP tasks, those simple features could be token-level features whose spurious association with the label can be learned easily. We show that MLM pretraining makes learning from the context easier. Thus, pretrained models are less likely to rely excessively on a single token. We also explore the theoretical explanations of MLM’s efficacy in causal settings. Compared with Wei et al. (2021), we achieve similar results with milder assumptions. Finally, we close the gap between our theories and real-world practices by conducting experiments on real-world tasks.

****

FLGAME: A Game-theoretic Defense against Backdoor Attacks In Federated Learning
Jinyuan Jia,Zhuowen Yuan,Dinuka Sahabandu,Luyao Niu,Arezoo Rajabi,Bhaskar Ramasubramanian,Bo Li,Radha Poovendran

Federated learning enables the distributed training paradigm, where multiple local clients jointly train a global model without needing to share their local training data. However, recent studies have shown that federated learning provides an additional surface for backdoor attacks. For instance, an attacker can compromise a subset of clients and thus corrupt the global model to incorrectly predict an attacker-chosen target class given any input embedded with the backdoor trigger. Existing defenses for federated learning against backdoor attacks usually detect and exclude the corrupted information from the compromised clients based on a $\textit{static}$ attacker model. Such defenses, however, are less effective when faced with $\textit{dynamic}$ attackers who can strategically adapt their attack strategies. In this work, we model the strategic interaction between the (global) defender and attacker as a minimax game. Based on the analysis of our model, we design an interactive defense mechanism that we call FLGAME. Theoretically, we prove that under mild assumptions, the global model trained with FLGAME under backdoor attacks is close to that trained without attacks. Empirically, we perform extensive evaluations on benchmark datasets and compare FLGAME with multiple state-of-the-art baselines. Our experimental results show that FLGAME can effectively defend against strategic attackers and achieves significantly higher robustness than baselines.

****

The Onset of Variance-Limited Behavior for Networks in the Lazy and Rich Regimes
Alexander Atanasov,Blake Bordelon,Sabarish Sainathan,Cengiz Pehlevan

For small training set sizes P, the generalization error of wide neural networks is well-approximated by the error of an infinite width neural network (NN), either in the kernel or mean-field/feature-learning regime. However, after a critical sample size P^*, we empirically find the finite-width network generalization becomes worse than that of the infinite width network. In this work, we empirically study the transition from infinite-width behavior to this \textit{variance-limited} regime as a function of sample size P and network width N. We find that finite-size effects can become relevant for very small dataset sizes on the order of $P^* \sim \sqrt{N}$ for polynomial regression with ReLU networks. We discuss the source of these effects using an argument based on the variance of the NN's final neural tangent kernel (NTK). This transition can be pushed to l arger P by enhancing feature learning or by ensemble averaging the networks. We find that the learning curve for regression with the final NTK is an accurate approximation of the NN learning curve. Using this, we provide a toy model which also exhibits $P^* \sim \sqrt{N}$ scaling and has P-dependent benefits from feature learning.

****

A Simple Approach for Visual Room Rearrangement: 3D Mapping and Semantic Search

Brandon Trabucco,Gunnar A Sigurdsson,Robinson Piramuthu,Gaurav S. Sukhatme,Ruslan Salakhutdinov

Physically rearranging objects is an important capability for embodied agents. Visual room rearrangement evaluates an agent's ability to rearrange objects in a room to a desired goal based solely on visual input. We propose a simple yet effective method for this problem: (1) search for and map which objects need to be rearranged, and (2) rearrange each object until the task is complete. Our approach consists of an off-the-shelf semantic segmentation model, voxel-based semantic map, and semantic search policy to efficiently find objects that need to be rearranged. Our method was the winning submission to the AI2-THOR Rearrangement Challenge in the 2022 Embodied AI Workshop at CVPR 2022, and improves on current state-of-the-art end-to-end reinforcement learning-based methods that learn visual room rearrangement policies from 0.53% correct rearrangement to 16.56%, using only 2.7% as many samples from the environment.

****

Memory Efficient Dynamic Sparse Training

Mike Heddes,Narayan Srinivasa

The excessive memory and energy consumption of modern Artificial Neural Networks (ANNs) is posing limitations on the machines that can run these models. Sparsification of ANNs is often motivated by time, memory and energy savings only during model inference, yielding no benefits during training. A growing body of work is now focusing on providing the benefits of model sparsification also during training. While these methods improve the energy efficiency during training, the algorithms yielding the most accurate models still have a peak memory usage on the same order as the dense model. We propose a Dynamic Sparse Training (DST) algorithm that reduces the peak memory usage during training while preserving the energy advantages of sparsely trained models. We evaluate our algorithm on CIFAR-10/100 using ResNet-56 and VGG-16 and compare it against a range of sparsification methods. The benefits of our method are twofold: first, it allows for a given model to be trained to an accuracy on par with the dense model while requiring significantly less memory and energy; second, the savings in memory and energy can be allocated towards training an even larger sparse model on the same machine, generally improving the accuracy of the model.

****

Accelerated Training via Principled Methods for Incrementally Growing Neural Networks

Xin Yuan,Pedro Henrique Pamplona Savarese,Michael Maire

We develop an approach to efficiently grow neural networks, within which parameterization and optimization strategies are designed by considering their effects on the training dynamics. Unlike existing growing methods, which follow simple replication heuristics or utilize auxiliary gradient-based local optimization, we craft a parameterization scheme which dynamically stabilizes weight, activation, and gradient scaling as the architecture evolves, and maintains the inference functionality of the network. To address the optimization difficulty resulting from imbalanced training effort distributed to subnetworks fading in at different growth phases, we propose a learning rate adaption mechanism that rebalances the gradient contribution of these separate subcomponents. Experimental results show that our method achieves comparable or better accuracy than training large fixed-size models, while saving a substantial portion of the original computation budget for training. We demonstrate that these gains translate into real wall-clock training speedups.

****

Progressive Mix-Up for Few-Shot Supervised Multi-Source Domain Transfer

Ronghang Zhu,Ronghang Zhu,Xiang Yu,Sheng Li

This paper targets at a new and challenging setting of knowledge transfer from multiple source domains to a single target domain, where target data is few shot or even one shot with label. Traditional domain generalization or adaptation methods cannot directly work since there is no sufficient target domain distribution serving as the transfer object. The multi-source setting further prevents the transfer task as excessive domain gap introduced from all the source domains. To tackle this problem, we newly propose a progressive mix-up (P-Mixup) mechanism to introduce an intermediate mix-up domain, pushing both the source domains and the few-shot target domain aligned to this mix-up domain. Further by enforcing the mix-up domain to progressively move towards the source domains, we achieve the domain transfer from multi-source domains to the single one-shot target domain. Our P-Mixup is different from traditional mix-up that ours is with a progressive and adaptive mix-up ratio, following the curriculum learning spirit to better align the source and target domains. Moreover, our P-Mixup combines both pixel-level and feature-level mix-up to better enrich the data diversity. Experiments on two benchmarks show that our P-Mixup significantly outperforms the state-of-the-art methods, i.e., 6.0\% and 6.8\% improvements on Office-Home and DomainNet.

****

Mitigating Propagation Failures in PINNs using Evolutionary Sampling

Arka Daw,Jie Bu,Sifan Wang,Paris Perdikaris,Anuj Karpatne

Despite the success of physics-informed neural networks (PINNs) in approximating partial differential equations (PDEs), it is known that PINNs can sometimes fail to converge to the correct solution in problems involving complicated PDEs. This is reflected in several recent studies on characterizing and mitigating the ``failure modes'' of PINNs. While most of these studies have focused on balancing loss functions or adaptively tuning PDE coefficients, what is missing is a thorough understanding of the connection between failure modes of PINNs and sampling strategies used for training PINNs. In this paper, we provide a novel perspective of failure modes of PINNs by hypothesizing that the training of PINNs rely on successful ``propagation'' of solution from initial and/or boundary condition points to interior points. We show that PINNs with poor sampling strategies can get stuck at trivial solutions if there are propagation failures. We additionally demonstrate that propagation failures are characterized by highly imbalanced PDE residual fields where very high residuals are observed over very narrow regions. To mitigate propagation failures, we propose a novel evolutionary sampling (Evo) method that can incrementally accumulate collocation points in regions of high PDE residuals with little to no computational overhead. We provide an extension of Evo to respect the principle of causality while solving time-dependent PDEs. We theoretically analyze the behavior of Evo and empirically demonstrate its efficacy and efficiency in comparison with baselines on a variety of PDE problems.

****

Revisiting Information-Based Clustering with Pseudo-Posterior Models

Zhongwen Zhang,Yuri Boykov

Maximization of mutual information (MI) between the network's input and output motivates standard losses for unsupervised discriminative clustering enforcing "decisiveness" and "fairness". In the context of common softmax models, we clarify several general properties of such discriminative losses that were previously not well understood: the relation to K-means, or lack thereof, and "margin-maximization". In particular, we show that "desiciveness" without the extra regularization term can lead to poor classification margins. Also, non-convexity of information-based losses motivates us to focus on self-supervised approaches introducing effective higher-order optimization algorithms with auxiliary variables. Addressing limitations of existing formulations, we propose a new self-supervised loss with soft auxiliary variables, or "pseudo-confidence" estimates. In particular, we introduce "strong" fairness and motivate the "reverse" cross-entropy as a robust loss for network training from noisy pseudo-confidence estimates. The latter is efficiently computed using variational inference - we derive a new EM algorithm with closed-form solutions for E and M steps. Empirically, our algorithm improves the performance of earlier methods for information-based clustering.

****

Neural Compositional Rule Learning for Knowledge Graph Reasoning

Kewei Cheng,Nesreen Ahmed,Yizhou Sun

Learning logical rules is critical to improving reasoning in KGs. This is due to their ability to provide logical and interpretable explanations when used for predictions, as well as their ability to generalize to other tasks, domains, and data. While recent methods have been proposed to learn logical rules, the majority of these methods are either restricted by their computational complexity and can not handle the large search space of large-scale KGs, or show poor generalization when exposed to data outside the training set. In this paper, we propose an end-to-end neural model for learning compositional logical rules called NCRL. NCRL detects the best compositional structure of a rule body, and breaks it into small compositions in order to infer the rule head. By recurrently merging compositions in the rule body with a recurrent attention unit, NCRL finally predicts a single rule head. Experimental results show that NCRL learns high-quality rules, as well as being generalizable. Specifically, we show that NCRL is scalable, efficient, and yields state-of-the-art results for knowledge graph completion on large-scale KGs. Moreover, we test NCRL for systematic generalization by learning to reason on small-scale observed graphs and evaluating on larger unseen ones.

****

Temporal Change Sensitive Representation for Reinforcement Learing

Qi Gao,Wei Xu

Image-based deep reinforcement learning has made a great improvement recently by combining state-of-the-art reinforcement learning algorithms with self-supervised representation learning algorithms. However, these self-supervised representation learning algorithms are designed to preserve global visual information, which may miss changes in visual information that are important for performing the task. To resolve this problem, self-supervised representation learning specifically designed for better preserving task relevant information is necessary. Following this idea, we introduce Temporal Change Sensitive Representation (TCSR), which is designed for reinforcement learning algorithms that have a latent dynamic model. TCSR enforces the latent state representation of the reinforcement agent to put more emphasis on the part of observation that could potentially change in the future. Our method achieves SoTA performance in Atari100K benchmark.

****

Provably Efficient Reinforcement Learning for Online Adaptive Influence Maximization

Kaixuan Huang,Yu Wu,Xuezhou Zhang,Shenyinying Tu,Qingyun Wu,Mengdi Wang,Huazheng Wang

Online influence maximization aims to maximize the influence spread of a content in a social network with an unknown network model by selecting a few seed nodes. Recent studies followed a non-adaptive setting, where the seed nodes are selected before the start of the diffusion process and network parameters are updated when the diffusion stops. We consider an adaptive version of content-dependent online influence maximization problem where the seed nodes are sequentially activated based on real-time feedback. In this paper, we formulate the problem as an infinite-horizon discounted MDP under a linear diffusion process and present a model-based reinforcement learning solution. Our algorithm maintains a network model estimate and selects seed users adaptively, exploring the social network while improving the optimal policy optimistically. We establish $\widetilde O(\sqrt{T})$ regret bound for our algorithm. Empirical evaluations on synthetic and real-world networks demonstrate the efficiency of our algorithm.

****

Efficient approximation of neural population structure and correlations with probabilistic circuits

Koosha Khalvati,Samantha Johnson,Stefan Mihalas,Michael A Buice

We present a computationally efficient framework to model a wide range of population structures with high order correlations and a large number of neurons. Our method is based on a special type of Bayesian network that has linear inference time and is founded upon the concept of contextual independence. Moreover, we use an efficient architecture learning method for network selection to model large neural populations even with a small amount of data. Our framework is both fast and accurate in approximating neural population structures. Furthermore, our approach enables us to reliably quantify higher order neural correlations. We test our method on simulated neural populations commonly used to generate higher order correlations, as well as on publicly available large-scale neural recordings from the Allen Brain Observatory. Our approach significantly outperforms other models both in terms of statistical measures and alignment with experimental evidence.

****

Exploring perceptual straightness in learned visual representations

Anne Harrington,Vasha DuTell,Ayush Tewari,Mark Hamilton,Simon Stent,Ruth Rosenholtz,William T. Freeman

Humans have been shown to use a ''straightened'' encoding to represent the natural visual world as it evolves in time (Henaff et al. 2019). In the context of discrete video sequences, ''straightened'' means that changes between frames follow a more linear path in representation space at progressively deeper levels of processing. While deep convolutional networks are often proposed as models of human visual processing, many do not straighten natural videos. In this paper, we explore the relationship between network architecture, differing types of robustness, biologically-inspired filtering mechanisms, and representational straightness in response to time-varying input; we identify strengths and limitations of straightness as a useful way of evaluating neural network representations. We find that (1) adversarial training leads to straighter representations in both CNN and transformer-based architectures but (2) this effect is task-dependent, not generalizing to tasks such as segmentation and frame-prediction, where straight representations are not favorable for predictions; and nor to other types of robustness. In addition, (3) straighter representations impart temporal stability to class predictions, even for out-of-distribution data. Finally, (4) biologically-inspired elements increase straightness in the early stages of a network, but do not guarantee increased straightness in downstream layers of CNNs. We show that straightness is an easily computed measure of representational robustness and stability, as well as a hallmark of human representations with benefits for computer vision models.

****

Improving Subgraph Representation Learning via Multi-View Augmentation

Yili Shen,Cheng-Wei Ju,Jiaxu Yan,Jun Yi,Zhou Lin,Hui Guan

Subgraph representation learning based on Graph Neural Network (GNN) has exhibited broad applications in scientific advancements, such as predictions of molecular structure-property relationships and collective cellular function. In particular, graph augmentation techniques have shown promising results in improving graph-based and node-based classification tasks. Still, they have rarely been explored in the existing GNN-based subgraph representation learning studies. In this study, we develop a novel multi-view augmentation mechanism to improve subgraph representation learning models and thus the accuracy of downstream prediction tasks. Our augmentation technique creates multiple variants of subgraphs and embeds these variants into the original graph to achieve highly improved training efficiency, scalability, and accuracy. Benchmark experiments on several real-world biological and physiological datasets demonstrate the superiority of our proposed multi-view augmentation techniques in subgraph representation learning.

****

System identification of neural systems: If we got it right, would we know?

Yena Han,Tomaso Poggio,Brian Cheung

Various artificial neural networks developed by engineers are now proposed as models of parts of the brain, such as the ventral stream in the primate visual cortex. The network activations are compared to recordings of biological neurons, and good performance in reproducing neural responses is considered to support the model's validity. This system identification approach, however, is only part of the traditional ways to develop and test models in the natural sciences. A key question is how much the ability to predict neural responses tells us. In particular, do these functional tests about neuron activation allow us to distinguish between different model architectures? We benchmark existing techniques to correctly identify a model by replacing brain recordings with known ground truth models. We evaluate the most commonly used identification approaches, such as a linear encoding model and centered kernel alignment. Even in the setting where the correct model is among the candidates, system identification performance is quite variable; it also depends significantly on factors independent of the ground truth architecture, such as stimuli images. In addition, we show the limitations of using functional similarity scores in identifying higher-level architectural motifs.

****

Is Forgetting Less a Good Inductive Bias for Forward Transfer?
Jiefeng Chen,Timothy Nguyen,Dilan Gorur,Arslan Chaudhry

One of the main motivations of studying continual learning is that the problem setting allows a model to accrue knowledge from past tasks to learn new tasks more efficiently. However, recent studies suggest that the key metric that continual learning algorithms optimize, reduction in catastrophic forgetting, does not correlate well with the forward transfer of knowledge. We believe that the conclusion previous works reached is due to the way they measure forward transfer. We argue that the measure of forward transfer to a task should not be affected by the restrictions placed on the continual learner in order to preserve knowledge of previous tasks. Instead, forward transfer should be measured by how easy it is to learn a new task given a set of representations produced by continual learning on previous tasks. Under this notion of forward transfer, we evaluate different continual learning algorithms on a variety of image classification benchmarks. Our results indicate that less forgetful representations lead to a better forward transfer suggesting a strong correlation between retaining past information and learning efficiency on new tasks. Further, we found less forgetful representations to be more diverse and discriminative compared to their forgetful counterparts.

****

High-Precision Regressors for Particle Physics
Fady Bishara,Ayan Paul,Jennifer Dy

Monte Carlo simulations of physics processes at particle colliders like the Large Hadron Collider at CERN take up a major fraction of the computational budget. For some simulations, a single data point takes seconds, minutes, or even hours to compute from first principles. Since the necessary number of data points per simulation is on the order of 10^9 -- 10^{12}, machine learning regressors can be used in place of physics simulators to significantly reduce this computational burden. However, this task requires high precision regressors that can deliver data with relative errors less than 1\% or even 0.1\% over the entire domain of the function. In this paper, we develop optimal training strategies and tune various machine learning regressors to satisfy the high-precision requirement. We leverage symmetry arguments from particle physics to optimize the performance of the regressors. Inspired by ResNets, we design a Deep Neural Network with skip connections that outperform fully connected Deep Neural Networks. We find that at lower dimensions, boosted decision trees far outperform neural networks while at higher dimensions neural networks perform better. We show that these regressors can speed up simulations by a factor of 10^3 -- 10^6 over the first-principles computations currently used in Monte Carlo simulations. Additionally, using symmetry arguments derived from particle physics, we reduce the number of regressors necessary for each simulation by an order of magnitude. Our work can significantly reduce the training and storage burden of Monte Carlo simulations at current and future collider experiments.

****

Learning Structured Representations by Embedding Class Hierarchy
Siqi Zeng,Remi Tachet des Combes,Han Zhao

Existing models for learning representations in supervised classification problems are permutation invariant with respect to class labels. However, structured knowledge about the classes, such as hierarchical label structures, widely exists in many real-world datasets, e.g., the ImageNet and CIFAR benchmarks. How to learn representations that can preserve such structures among the classes remains an open problem. To approach this problem, given a tree of class hierarchy, we first define a tree metric between any pair of nodes in the tree to be the length of the shortest path connecting them. We then provide a method to learn the hierarchical relationship of class labels by approximately embedding the tree metric in the Euclidean space of features. More concretely, during supervised training, we propose to use the Cophenetic Correlation Coefficient (CPCC) as a regularizer for the cross-entropy loss to correlate the tree metric of classes and the Euclidean distance in the class-conditioned representations. Our proposed regularizer is computationally lightweight and easy to implement. Empirically, we demonstrate that this approach can help to learn more interpretable representations due to the preservation of the tree metric, and leads to better in-distribution generalization as well as under sub-population shifts over six real-world datasets.

****

Promptagator: Few-shot Dense Retrieval From 8 Examples

Zhuyun Dai,Vincent Y Zhao,Ji Ma,Yi Luan,Jianmo Ni,Jing Lu,Anton Bakalov,Kelvin Guu,Keith Hall,Ming-Wei Chang

Much recent research on information retrieval has focused on how to transfer from one task (typically with abundant supervised data) to various other retrieval tasks where supervision is limited, with the implicit assumption that it is possible to generalize from one task to all the rest. However, this overlooks the fact that there are many diverse and unique retrieval problems, each targeting different search intents, queries, and search domains. In this paper, we suggest to work on Few-shot Dense Retrieval, a setting where each task comes with a short description and a few examples. To address this, we introduce Prompt-based Query Generation forRetrieval (Promptagator): for each task, we feed the few-shot examples to a large language model (LLM) and prompt it to behave as a task-specific query generator. Using this, we can synthetically generate a large number of relevant queries for any document, yielding abundant data for training task-specific retrievers --- with no reliance on traditional resources such as Natural Questions (Kwiatkowskiet al., 2019) or MS MARCO (Nguyen et al., 2016). Surprisingly, Promptagator with only 8 annotated examples enables efficient dual encoder retrievers to outperform computationally more expensive models trained on MS MARCO such as ColBERT v2 (Santhanam et al., 2022) by more than 1.2 points nDCG@10 on average on 11 retrieval sets. Further training standard-size re-rankers using the same generated data yields another 5.0 points nDCG@10 improvement. Our studies show that synthetic query generation can be far more effective than previously observed, especially when a small amount of task-specific knowledge is given.

****

Brain-like representational straightening of natural movies in robust feedforward neural networks

Tahereh Toosi,Elias Issa

Representational straightening refers to a decrease in curvature of visual feature representations of a sequence of frames taken from natural movies. Prior work established straightening in neural representations of the primate primary visual cortex (V1) and perceptual straightening in human behavior as a hallmark of biological vision in contrast to artificial feedforward neural networks which did not demonstrate this phenomenon as they were not explicitly optimized to produce temporally predictable movie representations. Here, we show robustness to noise in the input image can produce representational straightening in feedforward neural networks. Both adversarial training (AT) and base classifiers for Random Smoothing (RS) induced remarkably straightened feature codes. Demonstrating their utility within the domain of natural movies, these codes could be inverted to generate intervening movie frames by linear interpolation in the feature space even though they were not trained on these trajectories. Demonstrating their biolo gical utility, we found that AT and RS training improved predictions of neural data in primate V1 over baseline models providing a parsimonious, bio-plausible mechanism -- noise in the sensory input stages -- for generating representations in early visual cortex. Finally, we compared the geometric properties of frame representations in these networks to better understand how they produced representations that mimicked the straightening phenomenon from biology. Overall, this work elucidating emergent properties of robust neural networks demonstrates that it is not necessary to utilize predictive objectives or train directly on natural movie statistics to achieve models supporting straightened movie representations similar to human perception that also predict V1 neural responses.

****

FunkNN: Neural Interpolation for Functional Generation

AmirEhsan Khorashadizadeh,Anadi Chaman,Valentin Debarnot,Ivan Dokmani■

Can we build continuous generative models which generalize across scales, can be evaluated at any coordinate, admit calculation of exact derivatives, and are conceptually simple? Existing MLP-based architectures generate worse samples than the grid-based generators with favorable convolutional inductive biases. Models that focus on generating images at different scales do better, but employ complex architectures not designed for continuous evaluation of images and derivatives.

We take a signal-processing perspective and treat continuous signal generation as interpolation from samples. Indeed, correctly sampled discrete images contain all information about the low spatial frequencies. The question is then how to extrapolate the spectrum in a data-driven way while meeting the above design criteria. Our answer is FunkNN---a novel convolutional network which learns how to reconstruct continuous images at arbitrary coordinates and can be applied to any image dataset. Combined with a discrete generative model it becomes a functional generator which can act as a prior in continuous ill-posed inverse problems. We show that FunkNN generates high-quality continuous images and exhibits strong out-of-distribution performance thanks to its patch-based design. We further showcase its performance in several stylized inverse problems with exact spatial derivatives.

****

A Framework for Comprehensive Evaluations of Graph Neural Network based Community Detection using Node Clustering

Will Leeney,Ryan McConville

Graph Neural Networks (GNNs) have shown promising performance across a number of tasks in recent years. Unsupervised community detection using GNNs involves the clustering of nodes of a graph given both the features of nodes as well as the structure of the graph, and has many applications to real world tasks from social networks to genomics. Unfortunately, there has been relatively little research using GNNs for commOunity detection, and even less that evaluates the systems rigorously and fairly. A comprehensive evaluation of the performance of GNNs requires an suitable environment within which they are evaluated. This is exacerbated by the fact that community detection is primarily an unsupervised task, and that (graph) neural networks are used which contain many hyperparameters, discovered by inconsistent procedures. We argue that there is currently a gap in the literature that establishes a sufficient benchmarking environment for the consistent evaluation of GNN based community detection, thereby impeding progress in this nascent field. In this work we propose and evaluate an environment for the consistent evaluation of neural community detection. With this we show the strong dependence of the performance to the experimental settings , thereby motivating the use of this framework to facilitate research into GNN based community detection.

****

CrystalBox: Efficient Model-Agnostic Explanations for Deep RL Controllers

Sagar Patel,Sangeetha Abdu Jyothi,Nina Narodytska

Practical adoption of Reinforcement Learning (RL) controllers is hindered by a lack of explainability. Particularly, in input-driven environments such as comput er systems where the state dynamics are affected by external processes, explainability can serve as a key towards increased real-world deployment of RL controllers. In this work, we propose a novel framework, CrystalBox, for generating black-box post-hoc explanations for RL controllers in input-driven environments. CrystalBox is built on the principle of separation between policy learning and explanation computation. As the explanations are generated completely outside the training loop, CrystalBox is generalizable to a large family of input-driven RL controllers.To generate explanations, CrystalBox combines the natural decomposability of reward functions in systems environments with the explanatory power of decomposed returns. CrystalBox predicts these decomposed future returns using on policy Q-function approximations. Our design leverages two complementary approaches for this computation: sampling- and learning-based methods. We evaluate CrystalBox with RL controllers in real-world settings and demonstrate that it generates high-fidelity explanations.

****

Label Propagation with Weak Supervision

Rattana Pukdee,Dylan Sam,Pradeep Kumar Ravikumar,Nina Balcan

Semi-supervised learning and weakly supervised learning are important paradigms that aim to reduce the growing demand for labeled data in current machine learning applications. In this paper, we introduce a novel analysis of the classical label propagation algorithm (LPA) (Zhu & Ghahramani, 2002) that moreover takes advantage of useful prior information, specifically probabilistic hypothesized labels on the unlabeled data. We provide an error bound that exploits both the local geometric properties of the underlying graph and the quality of the prior information. We also propose a framework to incorporate multiple sources of noisy information. In particular, we consider the setting of weak supervision, where our sources of information are weak labelers. We demonstrate the ability of our approach on multiple benchmark weakly supervised classification tasks, showing improvements upon existing semi-supervised and weakly supervised methods.

****

TypeT5: Seq2seq Type Inference using Static Analysis

Jiayi Wei,Greg Durrett,Isil Dillig

There has been growing interest in automatically predicting missing type annotations in programs written in Python and JavaScript. While prior methods have achieved impressive accuracy when predicting the most common types, they often perform poorly on rare or complex types. In this paper, we present a new type inference method that treats type prediction as a code infilling task by leveraging CodeT5, a state-of-the-art seq2seq pre-trained language model for code. Our method uses static analysis to construct dynamic contexts for each code element whose type signature is to be predicted by the model. We also propose an iterative decoding scheme that incorporates previous type predictions in the model's input context, allowing information exchange between related code elements. Our evaluation shows that the proposed approach, TypeT5, not only achieves a higher overall accuracy (particularly on rare and complex types) but also produces more coherent results with fewer type errors---while enabling easy user intervention.

****

Approximating any Function via Coreset for Radial Basis Functions: Towards Provable Data Subset Selection For Efficient Neural Networks training

Murad Tukan,Samson Zhou,Alaa Maalouf,Vladimir Braverman,Dan Feldman

Radial basis function neural networks (\emph{RBFNN}) are notoriously known for their capability to approximate any continuous function on a closed bounded set with arbitrary precision given enough hidden neurons. Coreset is a small weighted subset of an input set of items, that provably approximates their loss function for a given set of queries (models, classifiers, etc.). In this paper, we suggest the first coreset construction algorithm for \emph{RBFNNs}, i.e., a small weighted subset which approximates the loss of the input data on any radial basis function network and thus approximates any function defined by an \emph{RBFNN} on the big input data. This is done by constructing coresets for radial basis and Laplacian loss functions. We use our coreset to suggest a provable data subset s election algorithm for training deep neural networks, since our coreset approximates every function, it should approximate the gradient of each weight in a neural network as it is defined as a function on the input. Experimental results on function approximation and dataset subset selection on popular network architectures and data sets are presented, demonstrating the efficacy and accuracy of our coreset construction.

****

Axiomatic Explainer Locality With Optimal Transport

Joshua Bone,Aria Masoomi,Jennifer Dy

Explainability methods have been notoriously difficult to evaluate and compare. Because of this, practitioners are often left guessing as to which explainer they should use for their task. Locality is one critical property of explainers which grants insight into the diversity of produced explanations. In this paper, we define a set of axioms which align with natural intuition regarding globalness, the inverse of locality. We then introduce a novel measure of globalness, Wasserstein Globalness, which uses optimal transport to quantify how local or global a given explainer is. Finally, we provide theoretical results describing the sample complexity of Wasserstein Globalness, and experimentally demonstrate how globalness can be used to effectively compare explainers. These results illustrate connections between both explainer fidelity and explainer robustness.

****

Fine-Tuning Offline Policies With Optimistic Action Selection

Max Sobol Mark,Ali Ghadirzadeh,Xi Chen,Chelsea Finn

Offline reinforcement learning algorithms can train performant policies for hard tasks using previously-collected datasets. However, the quality of the offline dataset often limits the levels of performance possible. We consider the problem of improving offline policies through online fine-tuning. Offline RL requires a pessimistic training objective to mitigate distributional shift between the trained policy and the offline behavior policy, which will make the trained policy averse to picking novel actions. In contrast, online RL requires exploration, or optimism. Thus, fine-tuning online policies with the offline training objective is not ideal. Additionally, loosening the fine-tuning objective to allow for more exploration can potentially destroy the behaviors learned in the offline phase because of the sudden and significant change in the optimization objective. To mitigate this challenge, we propose a method to facilitate exploration during online fine-tuning that maintains the same training objective throughout both offline and online phases, while encouraging exploration. We accomplish this by changing the action-selection method to be more optimistic with respect to the Q-function. By choosing to take actions in the environment with higher expected Q-values, our method is able to explore and improve behaviors more efficiently, obtaining 56% more returns on average than the alternative approaches on several locomotion, navigation, and manipulation tasks.

****

Improving the Strength of Human-Like Models in Chess

Saumik Narayanan,Kassa Korley,Chien-Ju Ho,Siddhartha Sen

Designing AI systems that capture human-like behavior has attracted growing attention in applications where humans may want to learn from, or need to collaborate with, these AI systems. Many existing works in designing human-like AI have taken a supervised learning approach that learns from data of human behavior, with the goal of creating models that can accurately predict human behavior. While this approach has shown success in capturing human behavior at different skill levels and even identifying individual behavioral styles, it also suffers from the drawback of mimicking human mistakes. Moreover, existing models only capture a snapshot of human behavior, leaving the question of how to improve them---e.g., from one human skill level to a stronger one---largely unanswered. Using chess as an experimental domain, we investigate the question of teaching an existing human-like model to be stronger using a data-efficient curriculum, while maintaining the model's human similarity. To achieve this goal, we extend the concept of curriculum learning to settings with multiple labeling strategies, allowing us to vary both the curriculum (dataset) and the teacher (labeling strategy). We fi nd that the choice of teacher has a strong impact on both playing strength and human similarity; for example, a teacher that is too strong can be less effective at improving playing strength and degrade human similarity more rapidly. We also find that the choice of curriculum can impact these metrics, but to a smaller extent; for example, training on a curriculum of human mistakes provides only a marginal benefit over training on a random curriculum. Finally, we show that our strengthened models achieve human similarity on datasets corresponding to their strengthened level of play, suggesting that our curriculum training methodology is improving them in human-like steps.

****

AGRO: Adversarial discovery of error-prone Groups for Robust Optimization

Bhargavi Paranjape,Pradeep Dasigi,Vivek Srikumar,Luke Zettlemoyer,Hannaneh Hajishirzi

Models trained via empirical risk minimization (ERM) are known to rely on spurious correlations between labels and task-independent input features, resulting in poor generalization to distributional shifts. Group distributionally robust optimization (G-DRO) can alleviate this problem by minimizing the worst-case loss over a set of pre-defined groups over training data. G-DRO successfully improves performance of the worst group, where the correlation does not hold. However, G-DRO assumes that the spurious correlations and associated worst groups are known in advance, making it challenging to apply them to new tasks with potentially multiple unknown correlations. We propose AGRO---Adversarial Group discovery for Distributionally Robust Optimization---an end-to-end approach that jointly identifies error-prone groups and improves accuracy on them. AGRO equips G-DRO with an adversarial slicing model to find a group assignment for training examples which maximizes worst-case loss over the discovered groups. On the WILDS benchmark, AGRO results in 8\% higher model performance on average on known worst-groups, compared to prior group discovery approaches used with G-DRO. AGRO also improves out-of-distribution performance on SST2, QQP, and MS-COCO---datasets where potential spurious correlations are as yet uncharacterized. Human evaluation of ARGO groups shows that they contain well-defined, yet previously unstudied spurious correlations that lead to model errors.

****

Learning Multiobjective Program Through Online Learning

Chaosheng Dong,Yijia Wang,Bo Zeng

We investigate the problem of learning the parameters (i.e., objective functions or constraints) of a multiobjective decision making model, based on a set of sequentially arrived decisions. In particular, these decisions might not be exact and possibly carry measurement noise or are generated with the bounded rationality of decision makers. In this paper, we propose a general online learning framework to deal with this learning problem using inverse multiobjective optimization, and prove that this framework converges at a rate of $\mathcal{O}(1/\sqrt{T})$ under certain regularity conditions. More precisely, we develop two online learning algorithms with implicit update rules which can handle noisy data. Numerical results with both synthetic and real world datasets show that both algorithms can learn the parameters of a multiobjective program with great accuracy and are robust to noise.

****

Dichotomy of Control: Separating What You Can Control from What You Cannot

Sherry Yang,Dale Schuurmans,Pieter Abbeel,Ofir Nachum

Future- or return-conditioned supervised learning is an emerging paradigm for offline reinforcement learning (RL), in which the future outcome (i.e., return) associated with a sequence of actions in an offline dataset is used as input to a policy trained to imitate those same actions. While return-conditioning is at the heart of popular algorithms such as decision transformer (DT), these methods tend to perform poorly in highly stochastic environments, where an occasional high return associated with a sequence of actions may be due more to the randomness of the environment than to the actions themselves. Such situations can lead to a learned policy that is inconsistent with its conditioning inputs; i.e., using the policy – while conditioned on a specific desired return – to act in the envi ronment can lead to a distribution of real returns that is wildly different than desired. In this work, we propose the dichotomy of control (DoC), a future-conditioned supervised learning framework that separates mechanisms within a policy’s control (actions) from those outside of a policy’s control (environment stochasticity). We achieve this by conditioning the policy on a latent variable representation of the future and designing a mutual information constraint that removes any future information from the latent variable that is only due to randomness of the environment. Theoretically, we show that DoC yields policies that are consistent with their conditioning inputs, ensuring that conditioning a learned policy on a desired high-return future outcome will correctly induce high-return behavior. Empirically, we show that DoC is able to achieve significantly better performance than DT on environments with highly stochastic rewards (e.g., Bandit) and transitions (e.g., FrozenLake).

****

Progressive Knowledge Distillation: Constructing Ensembles for Efficient Inference

Don Dennis,Abhishek Shetty,Anish Prasad Sevekari,Kazuhito Koishida,Virginia Smith

Knowledge distillation is commonly used to compress an ensemble of models into a single model. In this work we study the problem of progressive distillation: Given a large, pretrained teacher model g, we seek to decompose the model into an ensemble of smaller, low-inference cost student models f_i. The resulting ensemble allows for flexibly tuning accuracy vs. inference cost, which can be useful for a multitude of applications in efficient inference. Our method, B-DISTIL, uses a boosting procedure that allows function composition based aggregation rules to construct expressive ensembles with similar performance as g using much smaller student models. We demonstrate the effectiveness of B-DISTIL by decomposing pretrained models across a variety of image, speech, and sensor datasets. Our method comes with strong theoretical guarantees in terms of convergence as well as generalization.

****

Efficient Approximations of Complete Interatomic Potentials for Crystal Property Prediction

Yuchao Lin,Keqiang Yan,Youzhi Luo,Yi Liu,Xiaoning Qian,Shuiwang Ji

We study the problem of crystal material property prediction. A crystal structure consists of a minimal unit cell that is repeated infinitely in 3D space. How to accurately represent such repetitive structures in machine learning models remains unresolved. Current methods construct graphs by establishing edges only between nearby nodes, thereby failing to faithfully capture infinite repeating patterns and distant interatomic interactions. In this work, we propose several innovations to overcome these limitations. First, we propose to model physics-principled interatomic potentials directly instead of only using distances as in existing methods. These potentials include the Coulomb potential, London dispersion potential, and Pauli repulsion potential. Second, we propose to model the complete set of potentials among all atoms, instead of only between nearby atoms as in prior methods. This is enabled by our approximations of infinite potential summations with provable error bounds. We further develop efficient algorithms to compute the approximations. Finally, we propose to incorporate our computations of complete interatomic potentials into message passing neural networks for representation learning. We perform experiments on the JARVIS and Materials Project benchmarks for evaluation. Results show that the use of complete interatomic potentials leads to consistent performance improvements with reasonable computational costs.

****

LogicDP: Creating Labels for Graph Data via Inductive Logic Programming

Yuan Yang,Faramarz Fekri,James Clayton Kerce,Ali Payani

Graph data, such as scene graphs and knowledge graphs, see wide use in AI systems. In real-world and large applications graph data are usually incomplete, motivating graph reasoning models for missing-fact or missing-relationship inference. While these models can achieve state-of-the-art performance, they require a lar ge amount of training data.

Recent years have witnessed the rising interest in label creation with data programming (DP) methods, which aim to generate training labels from heuristic labeling functions. However, existing methods typically focus on unstructured data and are not optimized for graphs. In this work, we propose LogicDP, a data programming framework for graph data. Unlike existing DP methods, (1) LogicDP utilizes the inductive logic programming (ILP) technique and automatically discovers the labeling functions from the graph data; (2) LogicDP employs a budget-aware framework to iteratively refine the functions by querying an oracle, which significantly reduces the human efforts in function creations. Experiments show that LogicDP achieves better data efficiency in both scene graph and knowledge graph reasoning tasks.
****

Simulating Environments for Evaluating Scarce Resource Allocation Policies
Jeroen Berrevoets,Alex Chan,Daniel Jarrett,Mihaela van der Schaar
Consider the sequential decision problem of allocating a limited supply of resources to a pool of potential recipients: This scarce resource allocation problem arises in a variety of settings characterized by "hard-to-make" tradeoffs– such as assigning organs to transplant patients, or rationing ventilators in overstretched ICUs. Assisting human judgement in these choices are dynamic allocation policies that prescribe how to match available assets to an evolving pool of beneficiaries– such as clinical guidelines that stipulate selection criteria on the basis of recipient and organ attributes. However, while such policies have received increasing attention in recent years, a key challenge lies in pre-deployment evaluation: How might allocation policies behave in the real world? In particular, in addition to conventional backtesting, it is crucial that policies be evaluated on a variety of possible scenarios and sensitivities– such as distributions of recipients and organs that may diverge from historic patterns. In this work, we present AllSim, an open-source framework for performing data-driven simulation of scarce resource allocation policies for pre-deployment evaluation. Simulation environments are modular (i.e. parameterized componentwise), learnable (i.e. on historical data), and customizable (i.e. to unseen conditions), and– upon interaction with a policy –outputs a dataset of simulated outcomes for analysis and benchmarking. Compared to existing work, we believe this approach takes a step towards more methodical evaluation of scarce resource allocation policies.
****

Learning to reason with relational abstractions
Andrew Joo Hun Nam,Mengye Ren,Chelsea Finn,James Lloyd McClelland
Large language models have recently shown promising progress in mathematical reasoning when fine-tuned with human-generated sequences walking through a sequence of solution steps. However, the solution sequences are not formally structured and the resulting model-generated sequences may not reflect the kind of systematic reasoning we might expect an expert human to produce.

In this paper, we study how to build stronger reasoning capability in language models using the idea of relational abstractions. We introduce new types of sequences that more explicitly provide an abstract characterization of the transitions through intermediate solution steps to the goal state. We found that models that are supplied with such sequences as prompts can solve tasks with a significantly higher accuracy, and models that are trained to produce such sequences solve problems better than those that are trained with previously used human-generated sequences and other baselines. Our work thus takes several steps toward elucidating and improving how language models perform on tasks requiring multi-step mathematical reasoning.
****

A Simple Approach for State-Action Abstraction using a Learned MDP Homomorphism
Augustine N. Mavor-Parker,Matthew James Sargent,Andrea Banino,Lewis Griffin,Caswell Barry
Animals are able to rapidly infer from limited experience when sets of state action pairs have equivalent reward and transition dynamics. On the other hand, mod

ern reinforcement learning systems must painstakingly learn through trial and error that sets of state action pairs are value equivalent---requiring an often prohibitively large amount of samples from their environment. MDP homomorphisms have been proposed that reduce the observed MDP of an environment to an abstract MDP, which can enable more sample efficient policy learning. Consequently, impressive improvements in sample efficiency have been achieved when a suitable MDP homomorphism can be constructed a priori---usually by exploiting a practioner's knowledge of environment symmetries. We propose a novel approach to constructing a homomorphism in discrete action spaces, which uses a partial model of environment dynamics to infer which state action pairs lead to the same state---reducing the size of the state-action space by a factor equal to the cardinality of the action space. We call this method equivalent effect abstraction. We demonstrate empirically that equivalent effect abstraction can improve sample efficiency in a model-free setting and planning efficiency for model based approaches.

****

RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank

Quentin Garrido,Randall Balestriero,Laurent Najman,Yann LeCun

Joint-Embedding Self Supervised Learning (JE-SSL) has seen a rapid development, with the emergence of many method variations and few principled guidelines that would help practitioners to successfully deploy those methods. The main reason for that pitfall actually comes from JE-SSL's core principle of not employing any input reconstruction. Without any visual clue, it becomes extremely cryptic to judge the quality of a learned representation without having access to a labelled dataset. We hope to correct those limitations by providing a single --theoretically motivated-- criterion that reflects the quality of learned JE-SSL representations: their effective rank. Albeit simple and computationally friendly, this method ---coined {\em RankMe}--- allows one to assess the performance of JE-SSL representations, even on different downstream datasets, without requiring any labels, training or parameters to tune. Through thorough empirical experiments involving hundreds of repeated training episodes, we demonstrate how RankMe can be used for hyperparameter selection with nearly no loss in final performance compared to the current selection method that involve dataset labels. We hope that RankMe will facilitate the use of JE-SSL in domains with little or no labeled data.

****

Revisiting Intrinsic Reward for Exploration in Procedurally Generated Environments

Kaixin Wang,Kuangqi Zhou,Bingyi Kang,Jiashi Feng,Shuicheng YAN

Exploration under sparse rewards remains a key challenge in deep reinforcement learning. Recently, studying exploration in procedurally-generated environments has drawn increasing attention. Existing works generally combine lifelong intrinsic rewards and episodic intrinsic rewards to encourage exploration. Though various lifelong and episodic intrinsic rewards have been proposed, the individual contributions of the two kinds of intrinsic rewards to improving exploration are barely investigated. To bridge this gap, we disentangle these two parts and conduct ablative experiments. We consider lifelong and episodic intrinsic rewards used in prior works, and compare the performance of all lifelong-episodic combinations on the commonly used MiniGrid benchmark. Experimental results show that only using episodic intrinsic rewards can match or surpass prior state-of-the-art methods. On the other hand, only using lifelong intrinsic rewards hardly makes progress in exploration. This demonstrates that episodic intrinsic reward is more crucial than lifelong one in boosting exploration. Moreover, we find through experimental analysis that the lifelong intrinsic reward does not accurately reflect the novelty of states, which explains why it does not help much in improving exploration.

****

Online Learning for Obstacle Avoidance

David Snyder,Wenhan Xia,Daniel Suo,Anirudha Majumdar,Elad Hazan

We approach the fundamental problem of obstacle avoidance for robotic systems vi a the lens of online learning. In contrast to prior work that either assumes worst-case realization of uncertainty in the environment or a given stochastic model of uncertainty, we propose a method that is efficient to implement and provably grants instance-optimality to perturbations of trajectories generated from an open-loop planner in the sense of minimizing worst-case regret. The resulting policy thus adapts online to realizations of uncertainty and provably compares well with the best obstacle avoidance policy in hindsight from a rich class of policies. The method is validated in simulation on a dynamical system environment and compared to baseline open-loop planning and robust Hamilton-Jacobi reachability techniques.

****

Transformer-based World Models Are Happy With 100k Interactions

Jan Robine,Marc Höftmann,Tobias Uelwer,Stefan Harmeling

Deep neural networks have been successful in many reinforcement learning settings. However, compared to human learners they are overly data hungry. To build a sample-efficient world model, we apply a transformer to real-world episodes in an autoregressive manner: not only the compact latent states and the taken actions but also the experienced or predicted rewards are fed into the transformer, so that it can attend flexibly to all three modalities at different time steps. The transformer allows our world model to access previous states directly, instead of viewing them through a compressed recurrent state. By utilizing the Transformer-XL architecture, it is able to learn long-term dependencies while staying computationally efficient. Our transformer-based world model (TWM) generates meaningful, new experience, which is used to train a policy that outperforms previous model-free and model-based reinforcement learning algorithms on the Atari 100k benchmark. Our code is available at https://github.com/jrobine/twm.

****

Can Neural Networks Learn Implicit Logic from Physical Reasoning?

Aaron Traylor,Roman Feiman,Ellie Pavlick

Despite the success of neural network models in a range of domains, it remains an open question whether they can learn to represent abstract logical operators such as negation and disjunction. We test the hypothesis that neural networks without inherent inductive biases for logical reasoning can acquire an implicit representation of negation and disjunction. Here, implicit refers to limited, domain-specific forms of these operators, and work in psychology suggests these operators may be a precursor (developmentally and evolutionarily) to the type of abstract, domain-general logic that is characteristic of adult humans. To test neural networks, we adapt a test designed to diagnose the presence of negation and disjunction in animals and pre-verbal children, which requires inferring the location of a hidden object using constraints of the physical environment as well as implicit logic: if a ball is hidden in A or B, and shown not to be in A, can the subject infer that it is in B? Our results show that, despite the neural networks learning to track objects behind occlusion, they are unable to generalize to a task that requires implicit logic. We further show that models are unable to generalize to the test task even when they are trained directly on a logically identical (though visually dissimilar) task. However, experiments using transfer learning reveal that the models do recognize structural similarity between tasks which invoke the same logical reasoning pattern, suggesting that some desirable abstractions are learned, even if they are not yet sufficient to pass established tests of logical reasoning.

****

Blockwise self-supervised learning with Barlow Twins

Shoaib Ahmed Siddiqui,David Krueger,Yann LeCun,Stephane Deny

Current state-of-the-art deep networks are all powered by backpropagation. In this paper, we explore alternatives to full backpropagation in the form of blockwise learning rules, leveraging the latest developments in self-supervised learning. Notably, we show that a blockwise pretraining procedure consisting of training independently the 4 main blocks of layers of a ResNet-50 with Barlow Twins loss function at each block performs almost as well as end-to-end backpropagation on ImageNet: a linear probe trained on top of our blockwise pretrained model obta ins a top-1 classification accuracy of 70.48\%, only 1.1\% below the accuracy of an end-to-end pretrained network (71.57\% accuracy). We perform extensive experiments to understand the impact of different components within our method and explore a variety of adaptations of self-supervised learning to the blockwise paradigm, building an exhaustive understanding of the critical avenues for scaling local learning rules to large networks, with implications ranging from hardware design to neuroscience.

****

DIGEST: FAST AND COMMUNICATION EFFICIENT DECENTRALIZED LEARNING WITH LOCAL UPDATES

Peyman Gholami,Hulya Seferoglu

Decentralized learning advocates the elimination of centralized parameter servers (aggregation points) for potentially better utilization of underlying resources, delay reduction, and resiliency against parameter server unavailability and catastrophic failures. Gossip based decentralized algorithms, where each node in a network has its own locally kept model on which it effectuates the learning by talking to its neighbors, received a lot of attention recently. Despite their potential, Gossip algorithms introduce huge communication costs. In this work, we show that nodes do not need to communicate as frequently as in Gossip for fast convergence; in fact, a sporadic exchange of a digest of a trained model is sufficient. Thus, we design a fast and communication-efficient decentralized learning mechanism; DIGEST by particularly focusing on stochastic gradient descent (SGD). DIGEST is a decentralized algorithm building on local-SGD algorithms, which are originally designed for communication efficient centralized learning. We show through analysis and experiments that DIGEST significantly reduces the communication cost without hurting convergence time for both iid and non-iid data.

****

Learning to Improve Code Efficiency

Binghong Chen,Daniel Tarlow,Kevin Swersky,Martin Maas,Pablo Heiber,Ashish V Naik,Milad Hashemi,Parthasarathy Ranganathan

Improvements in the performance of computing systems, driven by Moore’s Law, have transformed society. As such hardware-driven gains slow down, it becomes even more important for software developers to focus on performance and efficiency during development. While several studies have demonstrated the potential from such improved code efficiency (e.g., 2x better generational improvements compared to hardware), unlocking these gains in practice has been challenging. Reasoning about algorithmic complexity and the interaction of coding patterns on hardware can be challenging for the average programmer, especially when combined with pragmatic constraints around development velocity and multi-person development.

This paper seeks to address this problem. We analyze a large competitive programming dataset from the Google Code Jam competition and find that efficient code is indeed rare, with a 2x runtime difference between the median and the 90th percentile of solutions. We propose using machine learning to automatically provide prescriptive feedback in the form of hints, to guide programmers towards writing high-performance code. To automatically learn these hints from the dataset, we propose a novel discrete variational auto-encoder, where each discrete latent variable represents a different learned category of code-edit that increases performance. We show that this method represents the multi-modal space of code efficiency edits better than a sequence-to-sequence baseline and generates a distribution of more efficient solutions.

****

ESCHER: Eschewing Importance Sampling in Games by Computing a History Value Function to Estimate Regret

Stephen Marcus McAleer,Gabriele Farina,Marc Lanctot,Tuomas Sandholm

Recent techniques for approximating Nash equilibria in very large games leverage neural networks to learn approximately optimal policies (strategies). One promis- ing line of research uses neural networks to approximate counterfactual regret minimization (CFR) or its modern variants. DREAM, the only current CFR-based neural method that is model free and therefore scalable to very large games, trains a neural network on an estimated regret target that can have extremely high variance due to an importance sampling term inherited from Monte Carlo CFR (MCCFR). In this paper we propose an unbiased model-free method that does not require any importance sampling. Our method, ESCHER, is principled and is guaranteed to converge to an approximate Nash equilibrium with high probability. We show that the variance of the estimated regret of ESCHER is orders of magnitude lower than DREAM and other baselines. We then show that ESCHER outperforms the prior state of the art—DREAM and neural fictitious self play (NFSP)—on a number of games and the difference becomes dramatic as game size increases. In the very large game of dark chess, ESCHER is able to beat DREAM and NFSP in a head-to-head competition over 90% of the time.

****

On Achieving Optimal Adversarial Test Error

Justin D. Li,Matus Telgarsky

We first elucidate various fundamental properties of optimal adversarial predictors: the structure of optimal adversarial convex predictors in terms of optimal adversarial zero-one predictors, bounds relating the adversarial convex loss to the adversarial zero-one loss, and the fact that continuous predictors can get arbitrarily close to the optimal adversarial error for both convex and zero-one losses. Applying these results along with new Rademacher complexity bounds for adversarial training near initialization, we prove that for general data distributions and perturbation sets, adversarial training on shallow networks with early stopping and an idealized optimal adversary is able to achieve optimal adversarial test error. By contrast, prior theoretical work either considered specialized data distributions or only provided training error guarantees.

****

General Policy Evaluation and Improvement by Learning to Identify Few But Crucial States

Francesco Faccio,Aditya Ramesh,Vincent Herrmann,Jean Harb,Jürgen Schmidhuber

Learning to evaluate and improve policies is a core problem of Reinforcement Learning (RL). Traditional RL algorithms learn a value function defined for a single policy. A recently explored competitive alternative is to learn a single value function for many policies. Here we combine the actor-critic architecture of Parameter-Based Value Functions and the policy embedding of Policy Evaluation Networks to learn a single value function for evaluating (and thus helping to improve) any policy represented by a deep neural network (NN). The method yields competitive experimental results. In continuous control problems with infinitely many states, our value function minimizes its prediction error by simultaneously learning a small set of `probing states' and a mapping from actions produced in probing states to the policy's return. The method extracts crucial abstract knowledge about the environment in form of very few states sufficient to fully specify the behavior of many policies. A policy improves solely by changing actions in probing states, following the gradient of the value function's predictions. Surprisingly, it is possible to clone the behavior of a near-optimal policy in Swimmer-v3 and Hopper-v3 environments only by knowing how to act in 3 and 5 such learned states, respectively. Remarkably, our value function trained to evaluate NN policies is also invariant to changes of the policy architecture: we show that it allows for zero-shot learning of linear policies competitive with the best policy seen during training.

****

Serving Graph Compression for Graph Neural Networks

Si Si,Felix Yu,Ankit Singh Rawat,Cho-Jui Hsieh,Sanjiv Kumar

Serving a GNN model online is challenging --- in many applications when testing nodes are connected to training nodes, one has to propagate information from training nodes to testing nodes to achieve the best performance, and storing the wh ole training set (including training graph and node features) during inference stage is prohibitive for large-scale problems. In this paper, we study graph compression to reduce the storage requirement for GNN in serving. Given a GNN model to be served, we propose to construct a compressed graph with a smaller number of nodes. In serving time, one just needs to replace the original training set graph by this compressed graph, without the need of changing the actual GNN model and the forward pass. We carefully analyze the error in the forward pass and derive simple ways to construct the compressed graph to minimize the approximation error. Experimental results on semi-supervised node classification demonstrate that the proposed method can significantly reduce the serving space requirement for GNN inference.

****

Optimal Data Sampling for Training Neural Surrogates of Programs

Alex Renda,Yi Ding,Michael Carbin

Programmers and researchers are increasingly developing surrogates of programs, models of a subset of the observable behavior of a given program, to solve a variety of software development challenges. Programmers train surrogates from measurements of the behavior of a program on a dataset of input examples.

We present a methodology for optimally sampling datasets to train neural network based surrogates of programs. We first characterize the optimal proportion of data to sample from each path in a program based on the complexity of learning the path. We next provide a program analysis to determine the complexity of different paths in a program. We evaluate these results on a large-scale graphics program, demonstrating that theoretically optimal sampling results in empirical improvements in accuracy.

****

Towards Understanding GD with Hard and Conjugate Pseudo-labels for Test-Time Adaptation

Jun-Kun Wang,Andre Wibisono

We consider a setting that a model needs to adapt to a new domain under distribution shifts, given that only unlabeled test samples from the new domain are accessible at test time. A common idea in most of the related works is constructing pseudo-labels for the unlabeled test samples and applying gradient descent (GD) to a loss function with the pseudo-labels. Recently, Goyal et al. (2022) propose conjugate labels, which is a new kind of pseudo-labels for self-training at test time. They empirically show that the conjugate label outperforms other ways of pseudo-labeling on many domain adaptation benchmarks. However, provably showing that GD with conjugate labels learns a good classifier for test-time adaptation remains open. In this work, we aim at theoretically understanding GD with hard and conjugate labels for a binary classification problem. We show that for square loss, GD with conjugate labels converges to an ϵ-optimal predictor under a Gaussian model for any arbitrarily small ϵ, while GD with hard pseudo-labels fails in this task. We also analyze them under different loss functions for the update. Our results shed lights on understanding when and why GD with hard labels or conjugate labels works in test-time adaptation.

****

Achieving Communication-Efficient Policy Evaluation for Multi-Agent Reinforcement Learning: Local TD-Steps or Batching?

FNU Hairi,Zifan Zhang,Jia Liu

In many consensus-based actor-critic multi-agent reinforcement learning (MARL) strategies, one of the key components is the MARL policy evaluation (PE) problem, where a set of N agents work cooperatively to evaluate the value function of the global states under a given policy only through communicating with their neighbors.

In MARL-PE, a critical challenge is how to lower the communication complexity, which is defined as the rounds of communication between neighboring nodes in order to converge to some ϵ-stationary point.

To lower communication complexity in MARL-PE, there exist two ``natural'' ideas: i) using batching to reduce the variance of TD (temporal difference) errors, wh

ich in turn improves the convergence rate of MARL-PE; and ii) performing multiple local TD update steps between each consecutive rounds of communication, so as to reduce the communication frequency.
While the effectiveness of the batching approach has been verified and relatively well-understood, the validity of the local TD-steps approach remains unclear due to the potential ``agent-drift'' phenomenon resulted from various heterogeneity factors across agents.
This leads to an interesting open question in MARL-PE: *Does the local TD-steps approach really work and how does it perform in comparison to the batching approach?*
In this paper, we take the first attempt to answer this interesting and fundamental question.
Our theoretical analysis and experimental results confirm that allowing multiple local TD steps is indeed a valid approach in lowering the communication complexity of MARL-PE compared to vanilla consensus-based MARL-PE algorithms.
Specifically, the local TD steps between two consecutive communication rounds can be as large as
$\mathcal{O}(\sqrt{1/\epsilon}\log{(1/\epsilon)})$ in order to converge to an ϵ-stationary point of MARL-PE.
Theoretically, we show that in order to reach the optimal sample complexity up to a log factor, the communication complexity is $\mathcal{O}(\sqrt{1/\epsilon}\log{(1/\epsilon)})$, which is *considerably worse* than TD learning with batching, whose communication complexity is $\mathcal{O}(\log (1/\epsilon))$. However, the experimental results show that the allowing multiple steps can be as good as the batch approach.
****
Learning where and when to reason in neuro-symbolic inference
Cristina Cornelio,Jan Stuehmer,Shell Xu Hu,Timothy Hospedales
The integration of hard constraints on neural network outputs is a very desirable capability. This allows to instill trust in AI by guaranteeing the sanity of that neural network predictions with respect to domain knowledge. Recently, this topic has received a lot of attention. However, all the existing methods usually either impose the constraints in a "weak" form at training time, with no guarantees at inference, or fail to provide a general framework that supports different tasks and constraint types. We tackle this open problem from a neuro-symbolic perspective. Our pipeline enhances a conventional neural predictor with (1) a symbolic reasoning module capable of correcting structured prediction errors and (2) a neural attention module that learns to direct the reasoning effort to focus on potential prediction errors, while keeping other outputs unchanged. This framework provides an appealing trade-off between the efficiency of constraint-free neural inference and the prohibitive cost of exhaustive reasoning at inference time. We show that our method outperforms the state of the art on visual-Sudoku, and can also benefit visual scene graph prediction. Furthermore, it can improve the performance of existing neuro-symbolic systems that lack our explicit reasoning during inference.
****
Aging with GRACE: Lifelong Model Editing with Key-Value Adaptors
Thomas Hartvigsen,Swami Sankaranarayanan,Hamid Palangi,Yoon Kim,Marzyeh Ghassemi
Large language models often err during deployment, either due to non-representative training data or distribution shift in the test set. Recently, model editors have been proposed to fix errors by adjusting a pre-trained model's weights. So far, however, existing model editors fail when making sequential edits by quickly decaying a model's performance on its upstream data. Further, when editing deployed online models, they quickly forget how to fix previously-seen mistakes. We advance beyond these existing methods by proposing and studying a novel Lifelong Model Editing setting, where errors stream into a deployed model and we update the model to correct its predictions without influencing its predictions for unrelated inputs. Towards effective methods in this challenging setting, we propose with General Retrieval Adaptors for Continual Editing, or GRACE. GRACE is a new Key-Value framework that casts model editing as a codebook update problem. Th

e proposed approach edits selected model layers by caching activations that are queried using embeddings from the previous layer. The cached activations are trained to correct a model's predictions, treating future layers as a decoder. As edits stream in, the keys and values of a GRACE layer are updated while the model weights remain frozen, ensuring similar edits are treated similarly without altering the model's performance on unrelated instances. Experimentally, we show that \method substantially improves over recent model editors.
****

A VAE for Transformers with Nonparametric Variational Information Bottleneck
James Henderson,Fabio James Fehr
We propose a Variational AutoEncoder (VAE) for Transformers by developing a Variational Information Bottleneck (VIB) regulariser for Transformer embeddings. We formalise such attention-based representations as mixture distributions, and use Bayesian nonparametrics to develop a Nonparametric VIB (NVIB) for them. The variable number of mixture components supported by nonparametrics captures the variable number of vectors supported by attention, and exchangeable distributions from nonparametrics capture the permutation invariance of attention. Our Transformer VAE (NVAE) uses NVIB to regularise the information passing from the Transformer encoder to the Transformer decoder. Evaluations of a NVAE, trained on natural language text, demonstrate that NVIB can regularise the number of mixture components in the induced embedding whilst maintaining generation quality and reconstruction capacity.
****

Learning MLPs on Graphs: A Unified View of Effectiveness, Robustness, and Efficiency
Yijun Tian,Chuxu Zhang,Zhichun Guo,Xiangliang Zhang,Nitesh Chawla
While Graph Neural Networks (GNNs) have demonstrated their efficacy in dealing with non-Euclidean structural data, they are difficult to be deployed in real applications due to the scalability constraint imposed by the multi-hop data dependency. Existing methods attempt to address this scalability issue by training student multi-layer perceptrons (MLPs) exclusively on node content features using labels derived from the teacher GNNs. However, the trained MLPs are neither effective nor robust. In this paper, we ascribe the lack of effectiveness and robustness to three significant challenges: 1) the misalignment between content feature and label spaces, 2) the strict hard matching to teacher's output, and 3) the sensitivity to node feature noises. To address the challenges, we propose NOSMOG, a novel method to learn NOise-robust Structure-aware MLPs On Graphs, with remarkable effectiveness, robustness, and efficiency. Specifically, we first address the misalignment by complementing node content with position features to capture the graph structural information. We then design an innovative representational similarity distillation strategy to inject soft node similarities into MLPs. Finally, we introduce adversarial feature augmentation to ensure stable learning against feature noises. Extensive experiments and theoretical analyses demonstrate the superiority of NOSMOG by comparing it to GNNs and the state-of-the-art method in both transductive and inductive settings across seven datasets. Codes are available at https://github.com/meettyj/NOSMOG.
****

On The Specialization of Neural Modules
Devon Jarvis,Richard Klein,Benjamin Rosman,Andrew M Saxe
A number of machine learning models have been proposed with the goal of achieving systematic generalization: the ability to reason about new situations by combining aspects of previous experiences. These models leverage compositional architectures which aim to learn specialized modules dedicated to structures in a task that can be composed to solve novel problems with similar structures. While the compositionality of these architectures is guaranteed by design, the modules specializing is not. Here we theoretically study the ability of network modules to specialize to useful structures in a dataset and achieve systematic generalization. To this end we introduce a minimal space of datasets motivated by practical systematic generalization benchmarks. From this space of datasets we present a mathematical definition of systematicity and study the learning dynamics of line

ar neural modules when solving components of the task. Our results shed light on the difficulty of module specialization, what is required for modules to successfully specialize, and the necessity of modular architectures to achieve systematicity. Finally, we confirm that the theoretical results in our tractable setting generalize to more complex datasets and non-linear architectures.

****

HomoDistil: Homotopic Task-Agnostic Distillation of Pre-trained Transformers

Chen Liang,Haoming Jiang,Zheng Li,Xianfeng Tang,Bing Yin,Tuo Zhao

Knowledge distillation has been shown to be a powerful model compression approach to facilitate the deployment of pre-trained language models in practice. This paper focuses on task-agnostic distillation. It produces a compact pre-trained model that can be easily fine-tuned on various tasks with small computational costs and memory footprints. Despite the practical benefits, task-agnostic distillation is challenging. Since the teacher model has a significantly larger capacity and stronger representation power than the student model, it is very difficult for the student to produce predictions that match the teacher's over a massive amount of open-domain training data. Such a large prediction discrepancy often diminishes the benefits of knowledge distillation. To address this challenge, we propose Homotopic Distillation (HomoDistil), a novel task-agnostic distillation approach equipped with iterative pruning. Specifically, we initialize the student model from the teacher model, and iteratively prune the student's neurons until the target width is reached. Such an approach maintains a small discrepancy between the teacher's and student's predictions throughout the distillation process, which ensures the effectiveness of knowledge transfer. Extensive experiments demonstrate that HomoDistil achieves significant improvements on existing baselines. Our codes will be released.

****

Information-Theoretic Underpinnings of Generalization and Translation in Emergent Communication

Mycal Tucker,Julie Shah,Roger P. Levy,Noga Zaslavsky

Traditional emergent communication (EC) methods often fail to generalize to novel settings or align with representations of natural language. While these limitations may at first appear unrelated, in this work, we show how controlling the Information Bottleneck (IB) tradeoff between complexity and informativeness (a principle thought to guide human languages) helps to address both of these problems in EC. Specifically, we build on VQ-VIB, a recently proposed method for training EC agents while controlling the IB tradeoff, in addition to maximizing agents' utility. We find that increasing informativeness, which is a task-agnostic measure of how well a listener can reconstruct a speaker's meaning, allows EC agents to better generalize to novel settings and more challenging tasks. At the same time, in translation experiments between EC and English, we find that increasing EC informativeness only improves team performance up to a certain threshold, corresponding to the English informativeness-complexity tradeoff. Jointly, our results indicate the importance of training EC systems while controlling the informativeness-complexity tradeoff to simultaneously support improved self-play performance and human-agent interaction.

****

Optimal Transport-Based Supervised Graph Summarization

Sepideh Neshatfar,Abram Magner,Salimeh Yasaei Sekeh

Graph summarization is the problem of producing smaller graph representations of an input graph dataset, in such a way that
  the smaller ``compressed'' graphs capture relevant structural information for downstream tasks. One graph summarization
  method, recently proposed in Garg & Jaakkola (2019), formulates an optimal transport-based framework that allows prior information
  about node, edge, and attribute importance to be incorporated into the graph summarization process. We extend the optimal transport framework to a supervised graph summarization setting, wherein we seek to preserve relevant information about a class label. We first formulate the problem in terms of maximizing the mutual information between the summarized graph and the class label. We then prop

ose a method that incorporates mutual information estimates between random variables associated with sample graphs and class labels into the optimal transport compression framework from Garg & Jaakkola (2019). We empirically show performance improvements over the previous work by Garg & Jaakkola (2019), in terms of classification and compression on synthetic and real datasets. We then theoretically show limitations of the optimal transport approach: e.g., that it fails to satisfy a certain desirable information monotonicity property.

****

Contrastive Vision Transformer for Self-supervised Out-of-distribution Detection

Hengding Wang,Yuchen Lu,Xi Chen

Out-of-distribution (OOD) detection is a type of technique that aims to detect abnormal samples that don't belong to the distribution of training data (or in-distribution (ID) data). The technique has been applied to various image classification tasks to identify abnormal image samples for which the abnormality is caused by semantic shift (from different classes) or covariate shift (from different domains). However, disentangling OOD samples caused by different shifts remains a challenge in image OOD detection. This paper proposes Contrastive Vision Transformer (CVT), an attention-based contrastive learning model, for self-supervised OOD detection in image classification tasks. Specifically, vision transformer architecture is integrated as a feature extracting module under a contrastive learning framework. An empirical ensemble module is developed to extract representative ensemble features, from which a balance can be achieved between semantic and covariate OOD samples. The proposed CVT model is tested in various self-supervised OOD detection tasks, and our approach outperforms state-of-the-art methods by 5.5% AUROC on CIFAR-10 (ID) vs. CIFAR-100 (OOD), and by 10.7% AUROC on CIFAR-100 (ID) vs. CIFAR-10 (OOD).

****

Does the Half Adversarial Robustness Represent the Whole? It Depends... A Theoretical Perspective of Subnetwork Robustness

Jovon Craig,Joshua Andle,Theodore Stein Nowak,Salimeh Yasaei Sekeh

Adversarial robustness of deep neural networks has been studied extensively and can bring security against adversarial attacks/examples. However, adversarially robust training approaches require a training mechanism on the entire deep network which can come at the cost of efficiency and computational complexity such as runtime. As a pilot study, we develop in this paper a novel theoretical framework that aims to answer the question of how can we make a whole model robust to adversarial examples by making part of a model robust? Toward promoting subnetwork robustness, we propose for the first time a new concept of semirobustness, which indicates adversarial robustness of a part of the network. We provide a theoretical analysis to show that if a subnetwork is robust and highly correlated with the rest of the network, then the remaining layers are also guaranteed to be robust. To guide the empirical investigation of our theoretical findings, we implemented our method at multiple layer depths and across multiple common image classification datasets. Experiments demonstrate that our method, with sufficient dependency between subnetworks, successfully utilizes subnetwork robustness to match fully-robust models' performance across AlexNet, VGG16, and ResNet50 benchmarks, for attack types FGSM, I-FGSM, PGD, and C&W.

****

Using Both Demonstrations and Language Instructions to Efficiently Learn Robotic Tasks

Albert Yu,Ray Mooney

Demonstrations and natural language instructions are two common ways to specify and teach robots novel tasks. However, for many complex tasks, a demonstration or language instruction alone contains ambiguities, preventing tasks from being specified clearly. In such cases, a combination of both a demonstration and an instruction more concisely and effectively conveys the task to the robot than either modality alone. To instantiate this problem setting, we train a single multi-task policy on a few hundred challenging robotic pick-and-place tasks and propose DeL-TaCo (Joint Demo-Language Task Conditioning), a method for conditioning a robotic policy on task embeddings comprised of two components: a visual demonstration and a language instruction. By allowing these two modalities to mutually disambiguate and clarify each other during novel task specification, DeL-TaCo (1) substantially decreases the teacher effort needed to specify a new task and (2) achieves better generalization performance on novel objects and instructions over previous task-conditioning methods. To our knowledge, this is the first work to show that simultaneously conditioning a multi-task robotic manipulation policy on both demonstration and language embeddings improves sample efficiency and generalization over conditioning on either modality alone.

****

On the duality between contrastive and non-contrastive self-supervised learning

Quentin Garrido,Yubei Chen,Adrien Bardes,Laurent Najman,Yann LeCun

Recent approaches in self-supervised learning of image representations can be categorized into different families of methods and, in particular, can be divided into contrastive and non-contrastive approaches. While differences between the two families have been thoroughly discussed to motivate new approaches, we focus more on the theoretical similarities between them. By designing contrastive and covariance based non-contrastive criteria that can be related algebraically and shown to be equivalent under limited assumptions, we show how close those families can be. We further study popular methods and introduce variations of them, allowing us to relate this theoretical result to current practices and show the influence (or lack thereof) of design choices on downstream performance. Motivated by our equivalence result, we investigate the low performance of SimCLR and show how it can match VICReg's with careful hyperparameter tuning, improving significantly over known baselines. We also challenge the popular assumption that non-contrastive methods need large output dimensions. Our theoretical and quantitative results suggest that the numerical gaps between contrastive and non-contrastive methods in certain regimes can be closed given better network design choices and hyperparameter tuning. The evidence shows that unifying different SOTA methods is an important direction to build a better understanding of self-supervised learning.

****

Few-Shot Incremental Learning Using HyperTransformers

Max Vladymyrov,Andrey Zhmoginov,Mark Sandler

Incremental few-shot learning methods make it possible to learn without forgetting from multiple few-shot tasks arriving sequentially. In this work we approach this problem using the recently published HyperTransformer (HT): a hypernetwork that generates task-specific CNN weights directly from the support set. We propose to re-use these generated weights as an input to the HT for the next task of the continual-learning sequence. Thus, the HT uses the weights themselves as the representation of the previously learned tasks. This approach is different from most continual learning algorithms that typically rely on using replay buffers, weight regularization or task-dependent architectural changes. Instead, we show that the HT works akin to a recurrent model, relying on the weights from the previous task and a support set from a new task. We demonstrate that a single HT equipped with a prototypical loss is capable of learning and retaining knowledge about past tasks for two continual learning scenarios: incremental-task learning and incremental-class learning.

****

The Brainy Student: Scalable Unlearning by Selectively Disobeying the Teacher

Meghdad Kurmanji,Peter Triantafillou,Eleni Triantafillou

Deep machine unlearning is the problem of removing the influence of a cohort of data from the weights of a trained deep model. This challenge has enjoyed increasing attention recently, motivated to the widespread use of neural networks in applications involving user data: allowing users to exercise their `right to be forgotten' necessitates an effective unlearning algorithm. Deleting data from models is also of interest in practice for removing out-of-date examples, outliers or noisy labels. However, most previous unlearning methods consider simple scenarios where a theoretical treatment is possible. Consequently, not only do their guarantees not apply to deep neural networks, but they also scale poorly. In th

is paper, drawing inspiration from teacher-student methods, we propose a scalable deep unlearning method that breaks free of previous limiting assumptions. Our thorough empirical investigation reveals that our approach significantly improves upon previous methods in being by far the most consistent in achieving unlearning in a wide range of scenarios, while incurring only a minimal performance degradation, if any, and being significantly more scalable than previous methods.
****

FIGARO: Controllable Music Generation using Learned and Expert Features
Dimitri von Rütte,Luca Biggio,Yannic Kilcher,Thomas Hofmann
Recent symbolic music generative models have achieved significant improvements in the quality of the generated samples. Nevertheless, it remains hard for users to control the output in such a way that it matches their expectation. To address this limitation, high-level, human-interpretable conditioning is essential. In this work, we release FIGARO, a Transformer-based conditional model trained to generate symbolic music based on a sequence of high-level control codes. To this end, we propose description-to-sequence learning, which consists of automatically extracting fine-grained, human-interpretable features (the description) and training a sequence-to-sequence model to reconstruct the original sequence given only the description as input. FIGARO achieves state-of-the-art performance in multi-track symbolic music generation both in terms of style transfer and sample quality. We show that performance can be further improved by combining human-interpretable with learned features. Our extensive experimental evaluation shows that FIGARO is able to generate samples that closely adhere to the content of the input descriptions, even when they deviate significantly from the training distribution.
****

A Neural PDE Solver with Temporal Stencil Modeling
Zhiqing Sun,Yiming Yang,Shinjae Yoo
Numerical simulation of non-linear partial differential equations plays a crucial role in modeling physical science and engineering phenomena, such as weather, climate, and aerodynamics. Recent Machine Learning (ML) models trained on low-resolution spatio-temporal signals have shown new promises in capturing important dynamics in high-resolution signals, under the condition that the models can effectively recover the missing details. However, this study shows that significant information is often lost in the low-resolution down-sampled features. To address such issues, we propose a new approach, namely Temporal Stencil Modeling (TSM), which combines the strengths of advanced time-series sequence modeling (with the HiPPO features) and state-of-the-art neural PDE solvers (with learnable stencil modeling). TSM aims to recover the lost information from the PDE trajectories and can be regarded as a temporal generalization of classic finite volume methods such as WENO. Our experimental results show that TSM achieves the new state-of-the-art simulation accuracy for 2-D incompressible Navier-Stokes turbulent flows: it significantly outperforms the previously reported best results by 19.9% in terms of the highly-correlated duration time, and reduces the inference latency into 80%. We also show a strong generalization ability of the proposed method to various out-of-distribution turbulent flow settings.
****

The Right Losses for the Right Gains: Improving the Semantic Consistency of Deep Text-to-Image Generation with Distribution-Sensitive Losses
Mahmoud Ahmed,Omer Moussa,Ismail Shaheen,Mohamed O Abdelfattah,Amr Adel Abdalla,Marwan Eid,Hesham Mohamed Eraqi,Mohamed N. Moustafa
One of the major challenges in training deep neural networks for text-to-image generation is the significant linguistic discrepancy between ground-truth captions of each image in most popular datasets. The large difference in the choice of words in such captions results in synthesizing images that are semantically dissimilar to each other and to their ground-truth counterparts. Moreover, existing models either fail to generate the fine-grained details of the image or require a huge number of parameters that renders them inefficient for text-to-image synthesis. To fill this gap in the literature, we propose using the contrastive learning approach with a novel combination of two loss functions: fake-to-fake loss

to increase the semantic consistency between generated images of the same caption, and fake-to-real loss to reduce the gap between the distributions of real images and fake ones. We test this approach on two baseline models: SSAGAN and AttnGAN (with style blocks to enhance the fine-grained details of the images.) Results show that our approach improves the qualitative results on AttnGAN with style blocks on the CUB dataset. Additionally, on the challenging COCO dataset, our approach achieves competitive results against the state-of-the-art Lafite model, outperforms the FID scores of SSAGAN and DALL-E models by 44% and 66.83% respectively, yet with only around 1% of the model size and training data of the huge DALL-E model.

****

Selection Collider Bias in Large Language Models

Emily McMilin

In this paper we motivate the causal mechanisms behind sample selection induced collider bias (selection collider bias) that can cause Large Language Mod- els (LLMs) to learn unconditional dependence between entities that are unconditionally independent in the real world. We show that selection collider bias can become amplified in underspecified learning tasks, and although difficult to overcome, we describe a method to exploit the resulting spurious correlations for determination of when a model may be uncertain about its prediction. We demonstrate an uncertainty metric that matches human uncertainty in tasks with gender pronoun underspecification on an extended version of the Winogender Schemas evaluation set, and we provide online demos where users can evaluate spurious correlations and apply our uncertainty metric to their own texts and models. Finally, we generalize our approach to address a wider range of prediction tasks.

****

CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data

Mathieu Chevalley,Yusuf H Roohani,Arash Mehrjou,Jure Leskovec,Patrick Schwab

Mapping biological mechanisms in cellular systems is a fundamental step in early-stage drug discovery that serves to generate hypotheses on what disease-relevant molecular targets may effectively be modulated by pharmacological interventions. With the advent of high-throughput methods for measuring single-cell gene expression under genetic perturbations, we now have effective means for generating evidence for causal gene-gene interactions at scale. However, inferring graphical networks of the size typically encountered in real-world gene-gene interaction networks is difficult in terms of both achieving and evaluating faithfulness to the true underlying causal graph. Moreover, standardised benchmarks for comparing methods for causal discovery in perturbational single-cell data do not yet exist. Here, we introduce CausalBench - a comprehensive benchmark suite for evaluating network inference methods on large-scale perturbational single-cell gene expression data. CausalBench introduces several biologically meaningful performance metrics and operates on two large, curated and openly available benchmark data sets for evaluating methods on the inference of gene regulatory networks from single-cell data generated under perturbations. With real-world datasets consisting of over 200000 training samples under interventions, CausalBench could potentially help facilitate advances in causal network inference by providing what is - to the best of our knowledge - the largest openly available test bed for causal discovery from real-world perturbation data to date.

****

Language models are multilingual chain-of-thought reasoners

Freda Shi,Mirac Suzgun,Markus Freitag,Xuezhi Wang,Suraj Srivats,Soroush Vosoughi,Hyung Won Chung,Yi Tay,Sebastian Ruder,Denny Zhou,Dipanjan Das,Jason Wei

We evaluate the reasoning abilities of large language models in multilingual settings. We introduce the Multilingual Grade School Math (MGSM) benchmark, by manually translating 250 grade-school math problems from the GSM8K dataset (Cobbe et al., 2021) into ten typologically diverse languages. We find that the ability to solve MGSM problems via chain-of-thought prompting emerges with increasing model scale, and that models have strikingly strong multilingual reasoning abilities, even in underrepresented languages such as Bengali and Swahili. Finally, we s how that multilingual reasoning abilities of language models extend to other tasks such as commonsense reasoning and word-in-context semantic judgment. The MGSM benchmark is publicly available at AnonymousLink and the supplementary material.

****

DreamFusion: Text-to-3D using 2D Diffusion
Ben Poole,Ajay Jain,Jonathan T. Barron,Ben Mildenhall

Recent breakthroughs in text-to-image synthesis have been driven by diffusion models trained on billions of image-text pairs. Adapting this approach to 3D synthesis would require large-scale datasets of labeled 3D or multiview data and efficient architectures for denoising 3D data, neither of which currently exist. In this work, we circumvent these limitations by using a pretrained 2D text-to-image diffusion model to perform text-to-3D synthesis. We introduce a loss based on probability density distillation that enables the use of a 2D diffusion model as a prior for optimization of a parametric image generator. Using this loss in a DeepDream-like procedure, we optimize a randomly-initialized 3D model (a Neural Radiance Field, or NeRF) via gradient descent such that its 2D renderings from random angles achieve a low loss. The resulting 3D model of the given text can be viewed from any angle, relit by arbitrary illumination, or composited into any 3D environment. Our approach requires no 3D training data and no modifications to the image diffusion model, demonstrating the effectiveness of pretrained image diffusion models as priors.

****

Recitation-Augmented Language Models
Zhiqing Sun,Xuezhi Wang,Yi Tay,Yiming Yang,Denny Zhou

We propose a new paradigm to help Large Language Models (LLMs) generate more accurate factual knowledge without retrieving from an external corpus, called RECITation-augmented gEneration (RECITE). Different from retrieval-augmented language models that retrieve relevant documents before generating the outputs, given an input, RECITE first recites one or several relevant passages from LLMs’ own memory via sampling, and then produces the final answers. We show that RECITE is a powerful paradigm for knowledge-intensive NLP tasks. Specifically, we show that by utilizing recitation as the intermediate step, a recite-and-answer scheme can achieve new state-of-the-art performance in various closed-book question answering (CBQA) tasks. In experiments, we verify the effectiveness of RECITE on three pre-trained models (In-house LM, UL2, and OPT) and three CBQA tasks (Natural Questions, TriviaQA, and HotpotQA). Our code is available at "https://github.com/Edward-Sun/RECITE".

****

Continual Active Learning
Arnav Mohanty Das,Gantavya Bhatt,Megh Manoj Bhalerao,Vianne R. Gao,Rui Yang,Jeff Bilmes

While active learning (AL) improves the labeling efficiency of machine learning (by allowing models to query the labels of data samples), a major problem is that compute efficiency is decreased since models are typically retrained from scratch at each query round. In this work, we develop a new framework that circumvents this problem by biasing further training towards the recently labeled sets, thereby complementing existing work on AL acceleration. We employ existing and novel replay-based Continual Learning (CL) algorithms that are effective at quickly learning new samples without forgetting previously learned information, especially when data comes from a shifting or evolving distribution. We call this compute-efficient active learning paradigm $\textit{``Continual Active Learning" (CAL)}$. We demonstrate that standard AL with warm starting fails, both to accelerate training, and that naive fine-tuning suffers from catastrophic forgetting due to distribution shifts over query rounds. We then show CAL achieves significant speedups using a plethora of replay schemes that use model distillation, and that select diverse/uncertain points from the history, all while maintaining performance on par with standard AL. We conduct experiments across many data domains, including natural language, vision, medical imaging, and computational biolo

gy, each with very different neural architectures (Transformers/CNNs/MLPs). CAL consistently provides a 2-6x reduction in training time, thus showing its applicability across differing modalities.
****

KwikBucks: Correlation Clustering with Cheap-Weak and Expensive-Strong Signals
Sandeep Silwal,Sara Ahmadian,Andrew Nystrom,Andrew McCallum,Deepak Ramachandran,Seyed Mehran Kazemi

The unprecedented rate at which the sizes of machine learning (ML) models are growing necessitates novel approaches to enable efficient and scalable solutions. We contribute to this line of work by studying a novel version of the Budgeted Correlation Clustering problem (\bcc) where along with a limited number of queries to an expensive oracle for node similarities (e.g. a large ML model), we have unlimited access to a cheaper but less accurate second oracle. Our formulation is inspired by many practical scenarios where coarse approximations of the expensive similarity metric can be efficiently obtained via weaker models. We develop a theoretically motivated algorithm in this setting that leverages the cheap oracle to judiciously query the strong oracle while maintaining high clustering quality. We empirically demonstrate gains in query minimization and clustering metrics on a variety of datasets with diverse strong and cheap oracles. Most notably, we demonstrate a practical application in text clustering based on expensive cross-attention language models by showing that cheaper (but weaker) embedding-based models can be leveraged to substantially reduce the number of inference calls to the former.
****

Credible, Sealed-bid, Optimal Repeated Auctions With Differentiable Economics
Davidson Cheng,Yang Hong,Ian Miers,John P Dickerson,Michael Curry

Online advertisement auctions happen billions of times per day. Bidders in auctions strategize to improve their own utility, subject to published auctions' rules. Yet, bidders may not know that an auction has been run as promised. A credible auction is one in which bidders can trust the auctioneer to run its allocation and pricing mechanisms as promised. It is known that, assuming no communication between bidders, no credible, sealed-bid, and incentive compatible (aka ``truth-telling'' or otherwise truthful-participation-incentivizing) mechanism can exist. In reality, bidders can certainly communicate, so what happens if we relax this (typically unrealistic) constraint?

In this work, we propose a framework incorporating cryptography to allow computationally-efficient, credible, revenue-maximizing (aka ``optimal'') auctions in a repeated auction setting. Our contribution is two-fold: first, we introduce a protocol for running repeated auctions with a verification scheme, and we show such a protocol can eliminate the auctioneer's incentive to deviate while costing negligible additional computation. Secondly, we provide a method for training optimal auctions under uncertain bidder participation profiles, which generalizes our protocol to a much wider class of auctions in the online ad market. Our empirical results show strong support for both the theory and competency of the proposed method.
****

The Power of Feel-Good Thompson Sampling: A Unified Framework for Linear Bandits
Zhiyuan Fan,Quanquan Gu

Linear contextual bandit is one of the most popular models in online decision-making with bandit feedback. Prior work has studied different variants of this model, e.g., misspecified, non-stationary, and multi-task/life-long linear contextual bandits. However, there is no single framework that can unify the algorithm design and analysis for these variants. In this paper, we propose a unified framework for linear contextual bandits based on feel-good Thompson sampling (Zhang, 2021). The algorithm derived from our framework achieves nearly minimax optimal regret in various settings and resolves the respective open problem in each setting. Specifically, let d be the dimension of the context and T be the length of the horizon, our algorithm achieves an $\widetilde{\mathcal{O}}(d\sqrt{ST})$ regret bound for non-stationary linear bandits with at most S switches, $\wid etilde{\mathcal{O}}(d^{\frac{5}{6}} T^{\frac{2}{3}} P^{\frac{1}{3}})$ regret for non-stationary linear bandits with bounded path length P, and $\widetilde{\mathcal{O}}(d\sqrt{kT} + \sqrt{dkMT})$ regret for (generalized) lifelong linear bandits over M tasks that share an unknown representation of dimension k. We believe our framework will shed light on the design and analysis of other linear contextual bandit variants.

****

Two-Tailed Averaging: Anytime Adaptive Once-in-a-while Optimal Iterate Averaging for Stochastic Optimization

Gábor Melis

Tail averaging improves on Polyak averaging's non-asymptotic behaviour by excluding a number of leading iterates of stochastic optimization from its calculations.

In practice, with a finite number of optimization steps and a learning rate that cannot be annealed to zero, tail averaging can get much closer to a local minimum point of the training loss than either the individual iterates or the Polyak average.

However, the number of leading iterates to ignore is an important hyperparameter, and starting averaging too early or too late leads to inefficient use of resources or suboptimal solutions.

Setting this hyperparameter to improve generalization is even more difficult, especially in the presence of other hyperparameters and overfitting.

Furthermore, before averaging starts, the loss is only weakly informative of the final performance, which makes early stopping unreliable.

To alleviate these problems, we propose an anytime variant of tail averaging, that has no hyperparameters and approximates the optimal tail at all optimization steps.

Our algorithm is based on two running averages with adaptive lengths bounded in terms of the optimal tail length, one of which achieves approximate optimality with some regularity.

Requiring only the additional storage for two sets of weights and periodic evaluation of the loss, the proposed two-tailed averaging algorithm is a practical and widely applicable method for improving stochastic optimization.

****

Reward Design with Language Models

Minae Kwon,Sang Michael Xie,Kalesha Bullard,Dorsa Sadigh

Reward design in reinforcement learning (RL) is challenging since specifying human notions of desired behavior may be difficult via reward functions or require many expert demonstrations. Can we instead cheaply design rewards using a natural language interface? This paper explores how to simplify reward design by using a large language model (LLM) such as GPT-3 as a proxy reward function, where the user provides a textual prompt containing a few examples (few-shot) or a description (zero-shot) of desired behavior. Our approach leverages this proxy reward function in an RL framework. Specifically, users specify a prompt once at the beginning of training. During training, the LLM evaluates an RL agent's behavior against the desired behavior described by the prompt and outputs a corresponding reward signal. The RL agent then uses this reward to update its behavior. We evaluate whether our approach can train agents aligned with user objectives in the Ultimatum Game, matrix games, and the DealOrNoDeal negotiation task. In all three tasks, we show that RL agents trained with our framework are well-aligned with the user's objectives and outperforms RL agents trained with reward functions learned via supervised learning.

****

Calibrating the Rigged Lottery: Making All Tickets Reliable

Bowen Lei,Ruqi Zhang,Dongkuan Xu,Bani Mallick

Although sparse training has been successfully used in various deep learning tasks to save memory and reduce inference time, the reliability of the produced sparse models remains unexplored. Previous research has shown that deep neural networks tend to be over-confident, and we find that sparse training exacerbates thi s problem. Therefore, calibrating the sparse models is crucial for reliable prediction and decision making. In this paper, we propose a new sparse training method to produce sparse models with improved confidence calibration. In contrast to previous research that uses only one mask to control the sparse topology, our method utilizes two masks, including a deterministic mask and a random mask. The former efficiently searches and activates important weights by exploiting the magnitude of weights and gradients. While the latter brings better exploration and finds more appropriate weight values by random updates. Theoretically, we prove our method can be viewed as a hierarchical variational approximation of a probabilistic deep Gaussian process. Extensive experiments on multiple datasets, model architectures, and sparsities show that our method can reduce ECE values by up to 47.8\% and simultaneously maintain or even improve accuracy with only a slight increase in computational and storage burden.

****

Replay Buffer with Local Forgetting for Adaptive Deep Model-Based Reinforcement Learning

Ali Rahimi-Kalahroudi,Janarthanan Rajendran,Ida Momennejad,Harm van Seijen,Sarath Chandar

One of the key behavioral characteristics used in neuroscience to determine whether the subject of study---be it a rodent or a human---exhibits model-based learning is effective adaptation to local changes in the environment. In reinforcement learning, however, recent work has shown that modern deep model-based reinforcement-learning (MBRL) methods adapt poorly to such changes. An explanation for this mismatch is that MBRL methods are typically designed with sample-efficiency on a single task in mind and the requirements for effective adaptation are substantially higher, both in terms of the learned world model and the planning routine. One particularly challenging requirement is that the learned world model has to be sufficiently accurate throughout relevant parts of the state-space. This is challenging for deep-learning-based world models due to catastrophic forgetting. And while a replay buffer can mitigate the effects of catastrophic forgetting, the traditional first-in-first-out replay buffer precludes effective adaptation due to maintaining stale data. In this work, we show that a conceptually simple variation of this traditional replay buffer is able to overcome this limitation. By removing only samples from the buffer from the local neighbourhood of the newly observed samples, deep world models can be built that maintain their accuracy across the state-space, while also being able to effectively adapt to changes in the reward function. We demonstrate this by applying our replay-buffer variation to the classical Dyna method, as well as to recent methods such as PlaNet and DreamerV2, showing for the first time that deep model-based methods are able to achieve effective adaptation.

****

Contrastive Audio-Visual Masked Autoencoder

Yuan Gong,Andrew Rouditchenko,Alexander H. Liu,David Harwath,Leonid Karlinsky,Hilde Kuehne,James R. Glass

In this paper, we first extend the recent Masked Auto-Encoder (MAE) model from a single modality to audio-visual multi-modalities. Subsequently, we propose the Contrastive Audio-Visual Masked Auto-Encoder (CAV-MAE) by combining contrastive learning and masked data modeling, two major self-supervised learning frameworks, to learn a joint and coordinated audio-visual representation.

Our experiments show that the contrastive audio-visual correspondence learning objective not only enables the model to perform audio-visual retrieval tasks, but also helps the model learn a better joint representation. As a result, our fully self-supervised pretrained CAV-MAE achieves a new SOTA accuracy of 65.9% on VGGSound, and is comparable with the previous best supervised pretrained model on AudioSet in the audio-visual event classification task. Code and pretrained models are at https://github.com/yuangongnd/cav-mae.

****

The Asymmetric Maximum Margin Bias of Quasi-Homogeneous Neural Networks

Daniel Kunin,Atsushi Yamamura,Chao Ma,Surya Ganguli

In this work, we explore the maximum-margin bias of quasi-homogeneous neural net works trained with gradient flow on an exponential loss and past a point of separability. We introduce the class of quasi-homogeneous models, which is expressive enough to describe nearly all neural networks with homogeneous activations, even those with biases, residual connections, and normalization layers, while structured enough to enable geometric analysis of its gradient dynamics. Using this analysis, we generalize the existing results of maximum-margin bias for homogeneous networks to this richer class of models. We find that gradient flow implicitly favors a subset of the parameters, unlike in the case of a homogeneous model where all parameters are treated equally. We demonstrate through simple examples how this strong favoritism toward minimizing an asymmetric norm can degrade the robustness of quasi-homogeneous models. On the other hand, we conjecture that this norm-minimization discards, when possible, unnecessary higher-order parameters, reducing the model to a sparser parameterization. Lastly, by applying our theorem to sufficiently expressive neural networks with normalization layers, we reveal a universal mechanism behind the empirical phenomenon of Neural Collapse.

****

Soft Diffusion: Score Matching For General Corruptions

Giannis Daras,Mauricio Delbracio,Hossein Talebi,Alex Dimakis,Peyman Milanfar

We define a broader family of corruption processes that generalizes previously known diffusion models. To reverse these general diffusions, we propose a new objective called Soft Score Matching that provably learns the score function for any linear corruption process and yields state of the art results for CelebA. Soft Score Matching incorporates the degradation process in the network.

Our new loss trains the model to predict a clean image, that after corruption, matches the diffused observation.

We show that our objective learns the gradient of the likelihood under suitable regularity conditions for a family of corruption processes.

We further develop a principled way to select the corruption levels for general diffusion processes and a novel sampling method that we call Momentum Sampler.

We show experimentally that our framework works for general linear corruption processes, such as Gaussian blur and masking.

We achieve state-of-the-art FID score 1.85 on CelebA-64, outperforming all previous linear diffusion models.

We also show significant computational benefits compared to vanilla denoising diffusion.

****

Open-Vocabulary Panoptic Segmentation MaskCLIP

Zheng Ding,Jieke Wang,Zhuowen Tu

In this paper, we tackle an emerging computer vision task, open-vocabulary panoptic segmentation, that aims to perform panoptic segmentation (background semantic labeling + foreground instance segmentation) for arbitrary categories of text-based descriptions in inference time. We first build a baseline method by directly adopting pre-trained CLIP models without finetuning nor distillation. We then develop MaskCLIP, a Transformer-based approach with a Relative Mask Attention (RMA) module. The RMA is an encoder-only module that seamless integrates mask tokens with a pre-trained ViT CLIP model for semantic/instance segmentation and class prediction. MaskCLIP learns to efficiently and effectively utilize pre-trained dense/local CLIP features within the RMA that avoids the time-consuming student-teacher training process. We obtain encouraging results for open-vocabulary panoptic/instance segmentation and state-of-the-art results for semantic segmentation on ADE20K and PASCAL datasets. We show qualitative illustration for MaskCLIP with online custom categories.

****

Robust Federated Learning with Majority Adversaries via Projection-based Re-weighting

Xiaoyang Wang,Klara Nahrstedt,Oluwasanmi O Koyejo

Most robust aggregators for distributed or federated learning assume that adversarial clients are the minority in the system. In contrast, this paper considers the majority adversary setting. We first show that a filtering method using a few trusted clients can defend against many standard attacks. However, a new attac k called Mimic-Shift can circumvent simple filtering. To this end, we develop a re-weighting strategy that identifies and down-weights the potential adversaries under the majority adversary regime. We show that our aggregator converges to a neighborhood around the optimum under the Mimic-Shift attack. Empirical results further show that our aggregator achieves negligible accuracy loss with a majority of adversarial clients, outperforming strong baselines.

****

Double Wins: Boosting Accuracy and Efficiency of Graph Neural Networks by Reliable Knowledge Distillation

Qiaoyu Tan,Daochen Zha,Soo-Hyun Choi,Li Li,Rui Chen,Xia Hu

The recent breakthrough achieved by graph neural networks (GNNs) with few labeled data accelerates the pace of deploying GNNs on real-world applications. While several efforts have been made to scale GNNs training for large-scale graphs, GNNs still suffer from the scalability challenge of model inference, due to the graph dependency issue incurred by the message passing mechanism, therefore hindering its deployment in resource-constrained applications. A recent study~\citep{zhang2021graph} revealed that GNNs can be compressed to inference-friendly multi-layer perceptrons (MLPs), by training MLPs using the soft labels of labeled and unlabeled nodes from the teacher. However, blindly leveraging the soft labels of all unlabeled nodes may be suboptimal, since the teacher model would inevitably make wrong predictions. This intriguing observation motivates us to ask: \textit{Is it possible to train a stronger MLP student by making better use of the unlabeled data?}

This paper studies cross-model knowledge distillation - from GNN teacher to MLP student in a semi-supervised setting, showing their strong promise in achieving a ``sweet point'' in co-optimizing model accuracy and efficiency. Our proposed solution, dubbed \textit{Reliable Knowledge Distillation for MLP optimization} (\textbf{RKD-MLP}), is the first noise-aware knowledge distillation framework for GNNs distillation. Its core idea is to use a meta-policy to filter out those unreliable soft labels. To train the meta-policy, we design a reward-driven objective based on a meta-set and adopt policy gradient to optimize the expected reward. Then we apply the meta-policy to the unlabeled nodes and select the most reliable soft labels for distillation. Extensive experiments across various GNN backbones, on 7 small graphs and 2 large-scale datasets from the challenging Open Graph Benchmark, demonstrate the superiority of our proposal. Moreover, our RKD-MLP model shows good robustness w.r.t. graph topology and node feature noises. The code is available at \url{https://anonymous.4open.science/r/RKD-MLP-F2A6/}.

****

A Statistical Framework for Personalized Federated Learning and Estimation: Theory, Algorithms, and Privacy

Kaan Ozkara,Antonious M. Girgis,Deepesh Data,Suhas Diggavi

A distinguishing characteristic of federated learning is that the (local) client data could have statistical heterogeneity. This heterogeneity has motivated the design of personalized learning, where individual (personalized) models are trained, through collaboration. There have been various personalization methods proposed in literature, with seemingly very different forms and methods ranging from use of a single global model for local regularization and model interpolation, to use of multiple global models for personalized clustering, etc. In this work, we begin with a statistical framework that unifies several different algorithms as well as suggest new algorithms. We apply our framework to personalized estimation, and connect it to the classical empirical Bayes' methodology. We develop novel private personalized estimation under this framework. We then use our statistical framework to propose new personalized learning algorithms, including AdaPeD based on information-geometry regularization, which numerically outperforms several known algorithms. We develop privacy for personalized learning methods with guarantees for user-level privacy and composition. We numerically evaluate the performance as well as the privacy for both the estimation and learning problems, demonstrating the advantages of our proposed methods.

****

Invariant Aggregator for Defending against Federated Backdoor Attacks
Xiaoyang Wang,Dimitrios Dimitriadis,Oluwasanmi O Koyejo,Shruti Tople

Federated learning is gaining popularity as it enables training of high-utility models across several clients without directly sharing their private data. As a downside, the federated setting makes the model vulnerable to various adversarial attacks in the presence of malicious clients. Specifically, an adversary can perform backdoor attacks to control model predictions via poisoning the training dataset with a trigger. In this work, we propose a mitigation for backdoor attacks in a federated learning setup. Our solution forces the model optimization trajectory to focus on the invariant directions that are generally useful for utility and avoid selecting directions that favor few and possibly malicious clients. Concretely, we consider the sign consistency of the pseudo-gradient (the client update) as an estimation of the invariance. Following this, our approach performs dimension-wise filtering to remove pseudo-gradient elements with low sign consistency. Then, a robust mean estimator eliminates outliers among the remaining dimensions. Our theoretical analysis further shows the necessity of the defense combination and illustrates how our proposed solution defends the federated learning model. Empirical results on three datasets with different modalities and varying number of clients show that our approach mitigates backdoor attacks with a negligible cost on the model utility.

****

Laser: Latent Set Representations for 3D Generative Modeling
Pol Moreno,Adam R. Kosiorek,Heiko Strathmann,Daniel Zoran,Rosalia Galiazzi Schneider,Björn Winckler,Larisa Markeeva,Theophane Weber,Danilo Jimenez Rezende

NeRF provides unparalleled fidelity of novel view synthesis---rendering a 3D scene from an arbitrary viewpoint. NeRF requires training on a large number of views that fully cover a scene, which limits its applicability.

While these issues can be addressed by learning a prior over scenes in various forms, previous approaches have been either applied to overly simple scenes or struggling to render unobserved parts.

We introduce Laser-NV---a generative model which achieves high modelling capacity, and which is based on a set-valued latent representation modelled by normalizing flows.

Similarly to previous amortized approaches, Laser-NV learns structure from multiple scenes and is capable of fast, feed-forward inference from few views.

To encourage higher rendering fidelity and consistency with observed views, Laser-NV further incorporates a geometry-informed attention mechanism over the observed views.

Laser-NV further produces diverse and plausible completions of occluded parts of a scene while remaining consistent with observations.

Laser-NV shows state-of-the-art novel-view synthesis quality when evaluated on ShapeNet and on a novel simulated City dataset, which features high uncertainty in the unobserved regions of the scene.

****

Towards Efficient Gradient-Based Meta-Learning in Heterogenous Environments
Thomas Goerttler,Luis Müller,Klaus Obermayer

A challenging problem for machine learning is few-shot learning, as its models usually require many training samples. Since meta-learning models have strong fine-tuning capabilities for the distribution of tasks, many of them have been applied to few-shot learning. Model-agnostic meta-learning (MAML) is one of the most popular ones. Recent studies showed that MAML-trained models tend to reuse learned features and do not perform strong adaption, especially in the earlier layers. This paper presents an in-detail analysis of this phenomenon by analyzing MAML's components of different variants. Our results show an interesting relationship between the importance of fine-tuning earlier layers and the difference in the distribution between training and testing. As a result, we determine a fundamental weakness of existing MAML variants when the task distribution is heterogeneous, e.g., the numbers of classes do not match during testing and training. We propose a novel nonparametric version of MAML that overcomes these issues while s

till being able to perform cross-domain adaption.
****

Optimal Transport for Offline Imitation Learning

Yicheng Luo,zhengyao jiang,Samuel Cohen,Edward Grefenstette,Marc Peter Deisenroth

With the advent of large datasets, offline reinforcement learning is a promising framework for learning good decision-making policies without the need to interact with the real environment.

However, offline RL requires the dataset to be reward-annotated, which presents practical challenges when reward engineering is difficult or when obtaining reward annotations is labor-intensive.

In this paper, we introduce Optimal Transport Relabeling (OTR), an imitation learning algorithm that can automatically relabel offline data of mixed and unknown quality with rewards from a few good demonstrations. OTR's key idea is to use optimal transport to compute an optimal alignment between an unlabeled trajectory in the dataset and an expert demonstration to obtain a similarity measure that can be interpreted as a reward, which can then be used by an offline RL algorithm to learn the policy. OTR is easy to implement and computationally efficient. On D4RL benchmarks, we demonstrate that OTR with a single demonstration can consistently match the performance of offline RL with ground-truth rewards.

****

FedorAS: Federated Architecture Search under system heterogeneity

■ukasz Dudziak,Stefanos Laskaridis,Javier Fernandez-Marques

Federated learning (FL) has recently gained considerable attention due to its ability to learn on decentralised data while preserving client privacy. However, it also poses additional challenges related to the heterogeneity of the participating devices, both in terms of their computational capabilities and contributed data. Meanwhile, Neural Architecture Search (NAS) has been successfully used with centralised datasets, producing state-of-the-art results in constrained or unconstrained settings. However, such centralised datasets may not be always available for training. Most recent work at the intersection of NAS and FL attempts to alleviate this issue in a cross-silo federated setting, which assumes homogeneous compute environments with datacenter-grade hardware.

In this paper we explore the question of whether we can design architectures of different footprints in a cross-device federated setting, where the device landscape, availability and scale are very different. To this end, we design our system, FedorAS, to discover and train promising architectures in a resource-aware manner when dealing with devices of varying capabilities holding non-IID distributed data. We present empirical evidence of its effectiveness across different settings, spanning across three different modalities (vision, speech, text), and showcase its better performance compared to state-of-the-art federated solutions, while maintaining resource efficiency.
****

Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization

Rajkumar Ramamurthy,Prithviraj Ammanabrolu,Kianté Brantley,Jack Hessel,Rafet Sifa,Christian Bauckhage,Hannaneh Hajishirzi,Yejin Choi

We tackle the problem of aligning pre-trained large language models (LMs) with human preferences. If we view text generation as a sequential decision-making problem, reinforcement learning (RL) appears to be a natural conceptual framework. However, using RL for LM-based generation faces empirical challenges, including training instability due to the combinatorial action space, as well as a lack of open-source libraries and benchmarks customized for LM alignment. Thus, a question rises in the research community: is RL a practical paradigm for NLP?

To help answer this, we first introduce an open-source modular library, $RL4LMs$ (Reinforcement Learning for Language Models), for optimizing language generators with RL. The library consists of on-policy RL algorithms that can be used to train any encoder or encoder-decoder LM in the HuggingFace library (Wolf et al. 2

020) with an arbitrary reward function. Next, we present the $GRUE$ (General Reinforced-language Understanding Evaluation) benchmark, a set of 6 language generation tasks which are supervised not by target strings, but by reward functions which capture automated measures of human preference.GRUE is the first leaderboard-style evaluation of RL algorithms for NLP tasks. Finally, we introduce an easy-to-use, performant RL algorithm, $NLPO$ (Natural Language Policy Optimization)} that learns to effectively reduce the combinatorial action space in language generation. We show 1) that RL techniques are generally better than supervised methods at aligning LMs to human preferences; and 2) that NLPO exhibits greater stability and performance than previous policy gradient methods (e.g., PPO (Schulman et al. 2017)), based on both automatic and human evaluations.
****

Learning multi-scale local conditional probability models of images
Zahra Kadkhodaie,Florentin Guth,Stéphane Mallat,Eero P Simoncelli
Deep neural networks can learn powerful prior probability models for images, as evidenced by the high-quality generations obtained with recent score-based diffusion methods. But the means by which these networks capture complex global statistical structure, apparently without suffering from the curse of dimensionality, remain a mystery. To study this, we incorporate diffusion methods into a multi-scale decomposition, reducing dimensionality by assuming a stationary local Markov model for wavelet coefficients conditioned on coarser-scale coefficients. We instantiate this model using convolutional neural networks (CNNs) with local receptive fields, which enforce both the stationarity and Markov properties. Global structures are captured using a CNN with receptive fields covering the entire (but small) low-pass image. We test this model on a dataset of face images, which are highly non-stationary and contain large-scale geometric structures.
Remarkably, denoising, super-resolution, and image synthesis results all demonstrate that these structures can be captured with significantly smaller conditioning neighborhoods than required by a Markov model implemented in the pixel domain. Our results show that score estimation for large complex images can be reduced to low-dimensional Markov conditional models across scales, alleviating the curse of dimensionality.
****

Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions
Sitan Chen,Sinho Chewi,Jerry Li,Yuanzhi Li,Adil Salim,Anru Zhang
We provide theoretical convergence guarantees for score-based generative models (SGMs) such as denoising diffusion probabilistic models (DDPMs), which constitute the backbone of large-scale real-world generative models such as DALL$\cdot$E 2. Our main result is that, assuming accurate score estimates, such SGMs can efficiently sample from essentially any realistic data distribution. In contrast to prior works, our results (1) hold for an L^2-accurate score estimate (rather than L^∞-accurate); (2) do not require restrictive functional inequality conditions that preclude substantial non-log-concavity; (3) scale polynomially in all relevant problem parameters; and (4) match state-of-the-art complexity guarantees for discretization of the Langevin diffusion, provided that the score error is sufficiently small. We view this as strong theoretical justification for the empirical success of SGMs. We also examine SGMs based on the critically damped Langevin diffusion (CLD). Contrary to conventional wisdom, we provide evidence that the use of the CLD does *not* reduce the complexity of SGMs.
****

Online Continual Learning with Feedforward Adaptation
ABULIKEMU ABUDUWEILI,Changliu Liu
Recently deep learning has been widely used in time-series prediction tasks. Although a trained deep neural network model typically performs well on the training set, performance drop significantly in a test set under slight distribution shifts. This challenge motivates the adoption of online adaptation algorithms to update the prediction models in real-time to improve the prediction performance.
Existing online adaptation methods optimize the prediction model by feeding back the latest prediction error computed with respect to the latest observation. Ho

wever, feedback based approach is prone to forgetting past information.

In this work, we propose an online adaptation method with feedforward compensation, which uses critical data samples from a memory buffer, instead of the latest samples, to optimize the prediction model. We prove that the proposed feedforward approach has a smaller error bound than the feedback approach in slow time-varying systems. The experiments on several time-series prediction tasks show that the proposed feedforward adaptation outperforms conventional feedback adaptation by more than 10%. In addition, the proposed feedforward adaptation method is able to estimate an uncertainty bound of the prediction that is agnostic from specific optimizers, while existing feedback adaptation could not.

****

Mind the Privacy Budget: How Generative Models Spend their Privacy Budgets

Georgi Ganev,Kai Xu,Emiliano De Cristofaro

Numerous Differentially Private (DP) generative models have been presented that aim to produce synthetic data while minimizing privacy risks.

As there is no single model that works well in all settings, empirical analysis is needed to establish and optimize trade-offs vis-\`a-vis the intended use of the synthetic data.

In this paper, we identify and address several challenges in the empirical evaluation of such models.

First, we analyze the steps in which different algorithms ``spend'' their privacy budget.

We evaluate the effects on the performance of downstream tasks to identify problem settings they are most likely to be successful at.

Then, we experiment with increasingly wider and taller training sets with various features, decreasing privacy budgets, and different DP mechanisms and generative models.

Our empirical evaluation, performed on both graphical and deep generative models, sheds light on the distinctive features of different models/mechanisms that make them well-suited for different settings and tasks.

Graphical models distribute the privacy budget horizontally and cannot handle relatively wide datasets, while the performance on the task they were optimized for monotonically increases with more data.

Deep generative models spend their budget per iteration, and their behavior is less predictable with varying dataset dimensions, but could perform better if trained on more features.

Also, low levels of privacy ($\epsilon\geq100$) could help some models generalize, achieving better results than without applying DP.

****

Resource Efficient Self-Supervised Learning for Speech Recognition

Abhinav Mehrotra,Alberto Gil Couto Pimentel Ramos,Nicholas Donald Lane,Sourav Bhattacharya

Representation learning from sequential data using self-supervised learning (SSL) has proven to be a powerful technique and improved state-of-the-art (SOTA) results when fine tuned for various downstream tasks, including Automatic Speech Recognition (ASR). So far the success of SSL frameworks, e.g., Wav2Vec-2.0, for sequence-to-sequence (seq2seq) modeling is primarily carried out by masking intermediate features and then solving a contrastive task in an end-to-end manner. Although very successful, the overall training time (for example, days or weeks) and demanding resource requirements for achieving SOTA performance remain a significant barrier to further improving ASR solutions using such approaches. In this work we show that non-contrastive learning, such as an extension of the Barlow–Twins methodology, when applied to seq2seq SSL modeling improves convergence, while reducing training time. Our results show that Wav2Vec-2.0 architecture pre-training with a non-contrastive SSL approach reduces the GPU training hours by 2.3 times, compared to masking based SSL approaches, while achieving a significant improvement (i.e., up to 6% relative WER decrease) in the model performance for the ASR task. We further demonstrate that a combination of both masking based SSL and non-contrastive SSL improves the ASR performance, e.g., up to 12% relative

WER decrease, for all splits of LibriSpeech evaluation dataset.

****

Subsampling in Large Graphs Using Ricci Curvature

Shushan Wu,Huimin Cheng,Jiazhang Cai,Ping Ma,Wenxuan Zhong

In the past decades, many large graphs with millions of nodes have been collected/constructed. The high computational cost and significant visualization difficulty hinder the analysis of large graphs. To overcome the difficulties, researchers have developed many graph subsampling approaches to provide a rough sketch that preserves global properties. By selecting representative nodes, these graph subsampling methods can help researchers estimate the graph statistics, e.g., the number of communities, of the large graph from the subsample. However, the available subsampling methods, e.g., degree node sampler and random walk sampler, tend to leave out minority communities because nodes with high degrees are more likely to be sampled. To overcome the shortcomings of the existing methods, we are motivated to apply the community information hidden in the graph to the subsampling method. Though the community structure is unavailable, community structure information can be obtained by applying geometric methods to a graph. An analog of Ricci curvature in the manifold is defined for the graph, i.e., Ollivier Ricci curvature. Based on the asymptotic results about the within-community edge and between-community edge's OR curvature, we propose a subsampling algorithm based on our theoretical results, the Ollivier-Ricci curvature Gradient-based subsampling (ORG-sub) algorithm. The proposed ORG-sub algorithm has two main contributions: First, ORG-sub provides a rigorous theoretical guarantee that the probability of ORG-sub taking all communities into the final subgraph converges to one. Second, extensive experiments on synthetic and benchmark datasets demonstrate the advantages of our algorithm.

****

Membership Leakage in Pre-trained Language Models

Yuan Xin,Zheng Li,Ning Yu,Michael Backes,Yang Zhang

Pre-trained language models are becoming a dominating component in NLP domain and have achieved state-of-the-art in various downstream tasks. Recent research has shown that language models are vulnerable to privacy leakage of their training data, such as text extraction and membership leakage. However, existing works against NLP applications mainly focus on the privacy leakage of text generation and downstream classification, and the privacy leakage of pre-trained language models is largely unexplored. In this paper, we take the first step toward systematically auditing the privacy risks of pre-trained language models through the lens of membership leakage. In particular, we focus on membership leakage of pre-training data in the exposure of downstream models adapted from pre-trained language models. We conduct extensive experiments on a variety of pre-trained model architectures and different types of downstream tasks. Our empirical evaluations demonstrate that membership leakage of pre-trained language models exists even when only the downstream model output is exposed, thereby posing a more severe risk than previously thought. We further conduct sophisticated ablation studies to analyze the relationship between membership leakage of pre-trained models and the characteristic of downstream tasks, which can guide developers or researchers to be vigilant about the vulnerability of pre-trained language models. Lastly, we explore possible defenses against membership leakage of PLMs and propose two promising defenses based on empirical evaluations.

****

Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching

Donggyun Kim,Jinwoo Kim,Seongwoong Cho,Chong Luo,Seunghoon Hong

Dense prediction tasks are a fundamental class of problems in computer vision. As supervised methods suffer from high pixel-wise labeling cost, a few-shot learning solution that can learn any dense task from a few labeled images is desired. Yet, current few-shot learning methods target a restricted set of tasks such as semantic segmentation, presumably due to challenges in designing a general and unified model that is able to flexibly and efficiently adapt to arbitrary tasks of unseen semantics. We propose Visual Token Matching (VTM), a universal few-sho t learner for arbitrary dense prediction tasks. It employs non-parametric matching on patch-level embedded tokens of images and labels that encapsulates all tasks. Also, VTM flexibly adapts to any task with a tiny amount of task-specific parameters that modulate the matching algorithm. We implement VTM as a powerful hierarchical encoder-decoder architecture involving ViT backbones where token matching is performed at multiple feature hierarchies. We experiment VTM on a challenging variant of Taskonomy dataset and observe that it robustly few-shot learns various unseen dense prediction tasks. Surprisingly, it is competitive with fully supervised baselines using only 10 labeled examples of novel tasks (0.004% of full supervision) and sometimes outperforms using 0.1% of full supervision. Codes are available at https://github.com/GitGyun/visual_token_matching.

****

The Game of Hidden Rules: A New Challenge for Machine Learning

Eric Pulick,Shubham Kumar Bharti,Yiding Chen,Vladimir Menkov,Yonatan Dov Mintz,Paul Kantor,Vicki M. Bier

Systematic examination of learning tasks remains an important but understudied area of machine learning (ML) research. To date, most ML research has focused on measuring performance on new tasks or surpassing state of the art performance on existing tasks. These efforts are vital but do not explain why some tasks are more difficult than others. Understanding how task characteristics affect difficulty is critical to formalizing ML's strengths and limitations; a rigorous assessment of which types of tasks are well-suited to a specific algorithm and, conversely, which algorithms are well-suited to a specific task would mark an important step forward for the field. To assist researchers in this effort, we introduce a novel learning environment designed to study how task characteristics affect measured difficulty for the learner. This tool frames learning tasks as a ``board-clearing game,'' which we call the Game of Hidden Rules (GOHR). In each instance of the game, the researcher encodes a specific rule, unknown to the learner, that determines which moves are allowed at each state of the game. The learner must infer the rule through play. We detail the game's expressive rule syntax and show how it gives researchers granular control over learning tasks. We present sample rules, a sample ML algorithm, and methods to assess algorithm performance. Separately, we provide additional benchmark rules, a public leaderboard for performance on these rules, and documentation for installing and using the GOHR environment.

****

Graph schemas as abstractions for transfer learning, inference, and planning

J Swaroop Guntupalli,Rajkumar Vasudeva Raju,Shrinu Kushagra,Danny Sawyer,Ishan Deshpande,Guangyao Zhou,Miguel Lazaro-Gredilla,Dileep George

We propose schemas as a model for abstractions that can be used for rapid transfer learning, inference, and planning. Common structured representations of concepts and behaviors---schemas---have been proposed as a powerful way to encode abstractions. Latent graph learning is emerging as a new computational model of the hippocampus to explain map learning and transitive inference. We build on this work to show that learned latent graphs in these models have a slot structure---schemas---that allow for quick knowledge transfer across environments. In a new environment, an agent can rapidly learn new bindings between the sensory stream to multiple latent schemas and select the best fitting one to guide behavior. To evaluate these graph schemas, we use two previously published challenging tasks: the memory \& planning game and one-shot StreetLearn, that are designed to test rapid task solving in novel environments. Graph schemas can be learned in far fewer episodes than previous baselines, and can model and plan in a few steps in novel variations of these tasks. We further demonstrate learning, matching, and reusing graph schemas in navigation tasks in more challenging environments with aliased observations and size variations, and show how different schemas can be composed to model larger environments.

****

Conservative Bayesian Model-Based Value Expansion for Offline Policy Optimization

Jihwan Jeong,Xiaoyu Wang,Michael Gimelfarb,Hyunwoo Kim,Baher abdulhai,Scott Sann er

Offline reinforcement learning (RL) addresses the problem of learning a performant policy from a fixed batch of data collected by following some behavior policy. Model-based approaches are particularly appealing in the offline setting since they can extract more learning signals from the logged dataset by learning a model of the environment. However, the performance of existing model-based approaches falls short of model-free counterparts, due to the compounding of estimation errors in the learned model. Driven by this observation, we argue that it is critical for a model-based method to understand when to trust the model and when to rely on model-free estimates, and how to act conservatively w.r.t. both. To this end, we derive an elegant and simple methodology called conservative Bayesian model-based value expansion for offline policy optimization (CBOP), that trades off model-free and model-based estimates during the policy evaluation step according to their epistemic uncertainties, and facilitates conservatism by taking a lower bound on the Bayesian posterior value estimate. On the standard D4RL continuous control tasks, we find that our method significantly outperforms previous model-based approaches: e.g., MOPO by 116.4%, MOReL by 23.2% and COMBO by 23.7%. Further, CBOP achieves state-of-the-art performance on 11 out of 18 benchmark datasets while doing on par on the remaining datasets.

****

Beam Tree Recursive Cells

Jishnu Ray Chowdhury,Cornelia Caragea

Recursive Neural Networks (RvNNs) generalize Recurrent Neural Networks (RNNs) by allowing sequential composition in a more flexible order, typically, based on some tree structure. While initially user-annotated tree structures were used, in due time, several approaches were proposed to automatically induce tree-structures from raw text to guide the recursive compositions in RvNNs. In this paper, we present an approach called Beam Tree Recursive Cell (or BT-Cell) based on a simple yet overlooked backpropagation-friendly framework. BT-Cell applies beam search on easy-first parsing for simulating RvNNs with automatic structure-induction. Our results show that BT-Cell achieves near-perfect performance on several aspects of challenging structure-sensitive synthetic tasks like ListOps and also comparable performance in realistic data to other RvNN-based models. We further introduce and analyze several extensions of BT-Cell based on relaxations of the hard top-k operators in beam search. We evaluate the models in different out of distribution splits in both synthetic and realistic data. Additionally, we identify a previously unknown failure case for neural models in generalization to unseen number of arguments in ListOps. We will release our code.

****

The Ultimate Combo: Boosting Adversarial Example Transferability by Composing Data Augmentations

Zebin Yun,Achi-Or Weingarten,Eyal Ronen,Mahmood Sharif

Transferring adversarial examples from surrogate (ML) models to evade target models is a common method for evaluating adversarial robustness in black-box settings. Researchers have invested substantial efforts to enhance transferability. Chiefly, attacks leveraging data augmentation have been found to help adversarial examples generalize better from surrogate to target models. Still, prior work has explored a limited set of augmentation techniques and their composition. To fill the gap, we conducted a systematic, comprehensive study of how data augmentation affects transferability. Particularly, we explored ten augmentation techniques of six categories originally proposed to help ML models generalize to unseen benign samples, and assessed how they influence transferability, both when applied individually and when composed. Our extensive experiments with the ImageNet dataset showed that simple color-space augmentations (e.g., color to greyscale) outperform the state of the art when combined with standard augmentations, such as translation and scaling. Additionally, except for two methods that may harm transferability, we found that composing augmentation methods impacts transferability monotonically (i.e., more methods composed $\rightarrow$ $\ge$transferability)---the best composition we found significantly outperformed the state of the art (e.g., 95.6% vs. 90.9% average transferability from normally trained surrogat

es to other normally trained models). We provide intuitive, empirically supported explanations for why certain augmentations fail to improve transferability.
****

Scaling up and Stabilizing Differentiable Planning with Implicit Differentiation
Linfeng Zhao,Huazhe Xu,Lawson L.S. Wong

Differentiable planning promises end-to-end differentiability and adaptivity. However, an issue prevents it from scaling up to larger-scale problems: they need to differentiate through forward iteration layers to compute gradients, which couples forward computation and backpropagation and needs to balance forward planner performance and computational cost of the backward pass. To alleviate this issue, we propose to differentiate through the Bellman fixed-point equation to decouple forward and backward passes for Value Iteration Network and its variants, which enables constant backward cost (in planning horizon) and flexible forward budget and helps scale up to large tasks. We study the convergence stability, scalability, and efficiency of the proposed implicit version of VIN and its variants and demonstrate their superiorities on a range of planning tasks: 2D navigation, visual navigation, and 2-DOF manipulation in configuration space and workspace.
****

Improving Aspect Ratio Distribution Fairness in Detector Pretraining via Cooperating RPN’s
Weilin Zhang,Xiang Li,Yu-Xiong Wang,David Forsyth

Region proposal networks (RPN) are a key component of modern object detectors. An RPN identifies image boxes likely to contain objects, and so worth further investigation. An RPN false negative is unrecoverable, so the performance of an object detector can be significantly affected by RPN behavior, particularly in low-data regimes. The RPN for a few shot detector is trained on base classes. Our experiments demonstrate that, if the distribution of box aspect ratios for base classes is different from that for novel classes, errors caused by RPN failure to propose a good box become significant. This is predictable: for example, an RPN trained on base classes that are mostly square will tend to miss short wide boxes. It has not been noticed to date because the (relatively few) standard base/novel class splits on current datasets do not display this effect. But changing the base/novel split highlights the problem. We describe datasets where the distribution shift is severe using PASCAL VOC, COCO, and LVIS datasets.

We show that the effect can be mitigated by training multiple distinct but cooperating specialized RPNs. Each specializes in a different aspect ratio, but cooperation constraints reduce the extent to which the RPNs are tuned. This means that if a box is missed by one RPN, it has a good chance of being picked up by another. Experimental evaluation confirms this approach results in substantial improvements in performance on the ARShift benchmarks, while remaining comparable to SOTA on conventional splits. Our approach applies to any few-shot detector and consistently improves performance of detectors.
****

UNDERSTANDING THE ROLE OF POSITIONAL ENCODINGS IN SENTENCE REPRESENTATIONS
Lihu Chen,Gael Varoquaux,Fabian M. Suchanek

Positional encodings are used to inject word-order information into transformer-based language models. While they can significantly enhance the quality of sentence representations, their specific contribution to language models are not fully understood, especially given recent findings that building natural-language understanding from language models with positional encodings is insensitive to word order. In this work, we investigate the role of positional encodings systematically. (1) We uncover the core function of existing positional encodings is to symmetrically combine local units by identifying two common properties, Locality, and Symmetry. (2) We reveal that positional and contextual encodings play a distinct role in understanding sentences. (3) Based on these findings, we propose a simplified new method to inject positional information into such models. Empirical studies demonstrate that this method can improve the performance of the BERT-based model on 10 downstream tasks. We hope these new probing results and findings can shed light on how to design and inject positional encodings into languag e models.

****

Artificial Replay: A Meta-Algorithm for Harnessing Historical Data in Bandits

Sid Banerjee,Sean R. Sinclair,Milind Tambe,Lily Xu,Christina Yu

While standard bandit algorithms sometimes incur high regret, their performance can be greatly improved by "warm starting" with historical data. Unfortunately, how best to incorporate historical data is unclear: naively initializing reward estimates using all historical samples can suffer from spurious data and imbalanced data coverage, leading to computational and storage issues - particularly in continuous action spaces. We address these two challenges by proposing Artificial Replay, a meta-algorithm for incorporating historical data into any arbitrary base bandit algorithm. Artificial Replayuses only a subset of the historical data as needed to reduce computation and storage. We show that for a broad class of base algorithms that satisfy independence of irrelevant data (IIData), a novel property that we introduce, our method achieves equal regret as a full warm-start approach while potentially using only a fraction of historical data. We complement these theoretical results with a case study of K-armed and continuous combinatorial bandit algorithms, including on a green security domain using real poaching data, to show the practical benefits of Artificial Replayin achieving optimal regret alongside low computational and storage costs.

****

Score-based Continuous-time Discrete Diffusion Models

Haoran Sun,Lijun Yu,Bo Dai,Dale Schuurmans,Hanjun Dai

Score-based modeling through stochastic differential equations (SDEs) has provided a new perspective on diffusion models, and demonstrated superior performance on continuous data. However, the gradient of the log-likelihood function, \ie, the score function, is not properly defined for discrete spaces. This makes it non-trivial to adapt SDE with score functions to categorical data. In this paper, we extend diffusion models to discrete variables by introducing a stochastic jump process where the reverse process denoises via a continuous-time Markov chain. This formulation admits an analytical simulation during backward sampling. To learn the reverse process, we extend score matching to general categorical data, and show that an unbiased estimator can be obtained via simple matching of the conditional marginal distributions. We demonstrate the effectiveness of the proposed method on a set of synthetic and real-world music and image benchmarks.

****

Decision Transformer under Random Frame Dropping

Kaizhe Hu,Ray Chen Zheng,Yang Gao,Huazhe Xu

Controlling agents remotely with deep reinforcement learning~(DRL) in the real world is yet to come. One crucial stepping stone is to devise RL algorithms that are robust in the face of dropped information from corrupted communication or malfunctioning sensors. Typical RL methods usually require considerable online interaction data that are costly and unsafe to collect in the real world. Furthermore, when applying to the frame dropping scenarios, they perform unsatisfactorily even with moderate drop rates. To address these issues, we propose Decision Transformer under Random Frame Dropping~(DeFog), an offline RL algorithm that enables agents to act robustly in frame dropping scenarios without online interaction. DeFog first randomly masks out data in the offline datasets and explicitly adds the time span of frame dropping as inputs. After that, a finetuning stage on the same offline dataset with a higher mask rate would further boost the performance. Empirical results show that DeFog outperforms strong baselines under severe frame drop rates like 90\%, while maintaining similar returns under non-frame-dropping conditions in the regular MuJoCo control benchmarks and the Atari environments. Our approach offers a robust and deployable solution for controlling agents in real-world environments with limited or unreliable data.

****

Semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation

Robin Khatri,Pierre Machart,Stefan Bonn

Cell deconvolution is the estimation of cell type fractions and cell type-specific gene expression from mixed data with unknown composition. In biomedical research, cell deconvolution, which is a source separation task, is used to obtain mechanistic and diagnostic insights into human diseases. An unmet challenge in cell deconvolution, however, is the scarcity of realistic training data and the strong domain shift observed in synthetic training data that is used in contemporary methods. Here, we hypothesize that simultaneous consistency regularization of the target and training domains will improve deconvolution performance. By adding this biologically motivated consistency loss to two novel deep learning-based deconvolution algorithms, we achieve state-of-the-art performance on both cell fraction and gene expression estimation. Our method, DISSECT, outperforms competing algorithms across several biomedical gene expression datasets and can be easily adapted to deconvolve other biomedical data types, as exemplified by our spatial expression deconvolution experiments.
****

Adversarial Imitation Learning with Preferences

Aleksandar Taranovic,Andras Gabor Kupcsik,Niklas Freymuth,Gerhard Neumann

Designing an accurate and explainable reward function for many Reinforcement Learning tasks is a cumbersome and tedious process.
Instead, learning policies directly from the feedback of human teachers naturally integrates human domain knowledge into the policy optimization process.
However, different feedback modalities, such as demonstrations and preferences, provide distinct benefits and disadvantages. For example, demonstrations convey a lot of information about the task but are often hard or costly to obtain from real experts while preferences typically contain less information but are in most cases cheap to generate.
However, existing methods centered around human feedback mostly focus on a single teaching modality, causing them to miss out on important training data while making them less intuitive to use.
In this paper we propose a novel method for policy learning that incorporates two different feedback types, namely \emph{demonstrations} and \emph{preferences}.

To this end, we make use of the connection between discriminator training and density ratio estimation to incorporate preferences into the popular Adversarial Imitation Learning paradigm.
This insight allows us to express loss functions over both demonstrations and preferences in a unified framework.
Besides expert demonstrations, we are also able to learn from imperfect ones and combine them with preferences to achieve improved task performance.
We experimentally validate the effectiveness of combining both preferences and demonstrations on common benchmarks and also show that our method can efficiently learn challenging robot manipulation tasks.
****

SuperFed: Weight Shared Federated Learning

Alind Khare,Animesh Agrawal,Alexey Tumanov

Federated Learning (FL) is a well-established technique for privacy preserving distributed training. Much attention has been given to various aspects of FL training. A growing number of applications that consume FL-trained models, however, increasingly operate under dynamically and unpredictably variable conditions, rendering a single model insufficient. We argue for training a global “family of models” cost efficiently in a federated fashion. Training them independently for different tradeoff points incurs ≈ O(k) cost for any k architectures of interest, however.
Straightforward applications of FL techniques to recent weight-shared training approaches is either infeasible or prohibitively expensive. We propose SuperFed — an architectural framework that incurs O(1) cost to co-train a large family of models in a federated fashion by leveraging weight-shared learning. We achieve an order of magnitude cost savings on both communication and computation by proposing two novel training mechanisms: (a) distribution of weight-shared models to federated clients, (b) central aggregation of arbitrarily overlapping weight-sha

red model parameters. The combination of these mechanisms is shown to reach an order of magnitude (9.43x) reduction in computation and communication cost for training a 5*10^18-sized family of models, compared to independently training as few as k = 9 DNNs without any accuracy loss.

****

Is Model Ensemble Necessary? Model-based RL via a Single Model with Lipschitz Regularized Value Function

Ruijie Zheng,Xiyao Wang,Huazhe Xu,Furong Huang

Probabilistic dynamics model ensemble is widely used in existing model-based reinforcement learning methods as it outperforms a single dynamics model in both asymptotic performance and sample efficiency. In this paper, we provide both practical and theoretical insights on the empirical success of the probabilistic dynamics model ensemble through the lens of Lipschitz continuity. We find that, for a value function, the stronger the Lipschitz condition is, the smaller the gap between the true dynamics- and learned dynamics-induced Bellman operators is, thus enabling the converged value function to be closer to the optimal value function. Hence, we hypothesize that the key functionality of the probabilistic dynamics model ensemble is to regularize the Lipschitz condition of the value function using generated samples. To validate this hypothesis, we devise two practical robust training mechanisms through computing the adversarial noise and regularizing the value network’s spectral norm to directly regularize the Lipschitz condition of the value functions. Empirical results show that combined with our mechanisms, model-based RL algorithms with a single dynamics model outperform those with ensemble of the probabilistic dynamics models. These findings not only support the theoretical insight, but also provide a practical solution for developing computationally efficient model-based RL algorithms.

****

Recurrent Back-Projection Generative Adversarial Network for Video Super Resolution

Israa Sabry Fahmy,Marwah Hisham Sulaiman,Zahraa Shehabeldin,Mohamed Elnaggar,Dareen Hussein Hasan,Mohamed Yasser,Moustafa Youssef,Hesham Mohamed Eraqi

In this paper, we propose a new Video Super Resolution algorithm in an attempt to generate videos that are temporally coherent, spatially detailed, and match human perception. To achieve this, we developed a new generative adversarial network named RBPGAN which is composed of two main components: a generator Network that exceeds other models for producing very high-quality frames, and a discriminator which outperforms others in terms of temporal consistency. The generator of the model uses a reduced recurrent back-projection network that takes a set of neighboring frames and a target frame applies SISR (Single Image Super Resolution) on each frame, and applies MISR (Multiple Image Super Resolution) through an encoder-decoder Back-Projection based approach to concatenate them and produce x4 resolution version of the target frame. The Spatio-temporal discriminator uses triplets of frames and penalizes the generator to generate the desired results. Our contribution results in a model that outperforms earlier work in terms of perceptual similarity and natural flow of frames, while maintaining temporal coherence and high-quality spatial details. The algorithm was tested on different datasets to eliminate bias.

****

Neural Networks as Paths through the Space of Representations

Richard D Lange,Devin Kwok,Jordan Kyle Matelsky,Xinyue Wang,David Rolnick,Konrad Kording

Deep neural networks implement a sequence of layer-by-layer operations that are each relatively easy to understand, but the resulting overall computation is generally difficult to understand. We develop a simple idea for interpreting the layer-by-layer construction of useful representations: the role of each layer is to reformat information to reduce the "distance" to the desired outputs. With this framework, the layer-wise computation implemented by a deep neural network can be viewed as a path through a high-dimensional representation space. We formalize this intuitive idea of a "path" by leveraging recent advances in metric representational similarity. We extend existing representational distance methods by computing geodesics, angles, and projections of representations, going beyond mere layer distances. We then demonstrate these tools by visualizing and comparing the paths taken by ResNet and VGG architectures on CIFAR-10. We conclude by sketching additional ways that this kind of representational geometry can be used to understand and interpret network training, to describe novel kinds of similarities between different models, and for representation-learning without backpropagation.

****

From Points to Functions: Infinite-dimensional Representations in Diffusion Models

Sarthak Mittal,Guillaume Lajoie,Stefan Bauer,Arash Mehrjou

Diffusion-based generative models learn to iteratively transfer unstructured noise to a complex target distribution as opposed to Generative Adversarial Networks (GANs) or the decoder of Variational Autoencoders (VAEs) which produce samples from the target distribution in a single step. Thus, in diffusion models every sample is naturally connected to a random trajectory which is a solution to a learned stochastic differential equation (SDE). Generative models are only concerned with the final state of this trajectory that delivers samples from the desired distribution. \cite{abstreiter2021diffusion} showed that these stochastic trajectories can be seen as continuous filters that wash out information along the way. Consequently, it is reasonable to ask if there is an intermediate time step at which the preserved information is optimal for a given downstream task. In this work, we show that a combination of information content from different time steps gives a strictly better representation for the downstream task. We introduce an attention and recurrence based modules that ``learn to mix'' information content of various time-steps such that the resultant representation leads to superior performance in downstream tasks.

****

Disentanglement with Biological Constraints: A Theory of Functional Cell Types

James C. R. Whittington,Will Dorrell,Surya Ganguli,Timothy Behrens

Neurons in the brain are often finely tuned for specific task variables. Moreover, such disentangled representations are highly sought after in machine learning. Here we mathematically prove that simple biological constraints on neurons, namely nonnegativity and energy efficiency in both activity and weights, promote such sought after disentangled representations by enforcing neurons to become selective for single factors of task variation. We demonstrate these constraints lead to disentanglement in a variety of tasks and architectures, including variational autoencoders. We also use this theory to explain why the brain partitions its cells into distinct cell types such as grid and object-vector cells, and also explain when the brain instead entangles representations in response to entangled task factors. Overall, this work provides a mathematical understanding of why single neurons in the brain often represent single human-interpretable factors, and steps towards an understanding task structure shapes the structure of brain representation.

****

Efficient One-Shot Neural Architecture Search With Progressive Choice Freezing Evolutionary Search

Chen Zhang,Qiyu Wan,Lening Wang,Mingsong Chen,Jingweijia Tan,Kaige Yan,Xin Fu

Neural Architecture Search (NAS) is a fast-developing research field to promote automatic machine learning. Among the recently populated NAS methods, One-Shot NAS has attracted significant attention since it greatly reduces the training cost compared with the previous NAS methods. In One-Shot NAS, the best network architecture is searched within a supernet, which is trained only once. In practice, the search process involves numerous inference processes for each user case, which causes high overhead in terms of latency and energy consumption. To tackle this problem, we first observe that the choices of the first few blocks that belong to different candidate networks will become similar at the early search stage. Furthermore, these choices are already close to the optimal choices obtained at the end of the search. Leveraging this interesting feature, we propose a Progr essive Choice Freezing Evolutionary Search (PCF-ES) method that gradually freezes block choices for all subnets at different search generations. This approach gives us an opportunity to reuse intermediate data produced by the frozen block instead of re-computing them. The experiment results show that the proposed PCF-ES provides up to 55\% speedup and reduces energy consumption by 51\% during the searching stage.

****

Synthetic Data Generation of Many-to-Many Datasets via Random Graph Generation

Kai Xu,Georgi Ganev,Emile Joubert,Rees Davison,Olivier Van Acker,Luke Robinson

Synthetic data generation (SDG) has become a popular approach to release private datasets.

In SDG, a generative model is fitted on the private real data, and samples drawn from the model are released as the protected synthetic data.

While real-world datasets usually consist of multiple tables with potential \emph{many-to-many} relationships (i.e.~\emph{many-to-many datasets}), recent research in SDG mostly focuses on modeling tables \emph{independently} or only considers generating datasets with special cases of many-to-many relationships such as \emph{one-to-many}.

In this paper, we first study challenges of building faithful generative models for many-to-many datasets, identifying limitations of existing methods.

We then present a novel factorization for many-to-many generative models, which leads to a scalable generation framework by combining recent results from random graph theory and representation learning.

Finally, we extend the framework to establish the notion of (ϵ,δ)-differential privacy.

Through a real-world dataset, we demonstrate that our method can generate synthetic datasets while preserving information within and across tables better than its closest competitor.

****

Learning rigid dynamics with face interaction graph networks

Kelsey R Allen,Yulia Rubanova,Tatiana Lopez-Guevara,William F Whitney,Alvaro Sanchez-Gonzalez,Peter Battaglia,Tobias Pfaff

Simulating rigid collisions among arbitrary shapes is notoriously difficult due to complex geometry and the strong non-linearity of the interactions. While graph neural network (GNN)-based models are effective at learning to simulate complex physical dynamics, such as fluids, cloth and articulated bodies, they have been less effective and efficient on rigid-body physics, except with very simple shapes. Existing methods that model collisions through the meshes' nodes are often inaccurate because they struggle when collisions occur on faces far from nodes. Alternative approaches that represent the geometry densely with many particles are prohibitively expensive for complex shapes. Here we introduce the ``Face Interaction Graph Network'' (FIGNet) which extends beyond GNN-based methods, and computes interactions between mesh faces, rather than nodes. Compared to learned node- and particle-based methods, FIGNet is around 4x more accurate in simulating complex shape interactions, while also 8x more computationally efficient on sparse, rigid meshes. Moreover, FIGNet can learn frictional dynamics directly from real-world data, and can be more accurate than analytical solvers given modest amounts of training data. FIGNet represents a key step forward in one of the few remaining physical domains which have seen little competition from learned simulators, and offers allied fields such as robotics, graphics and mechanical design a new tool for simulation and model-based planning.

****

On the Importance of Contrastive Loss in Multimodal Learning

Yunwei Ren,Yuanzhi Li

Recently, contrastive learning approaches (e.g., CLIP (Radford et al., 2021)) have received huge success in multimodal learning, where the model tries to minimize the distance between the representations of different views (e.g., image and its caption) of the same data point, while keep the representations of different data points away from each other. However, from a theoretical perspective, it is unclear how contrastive learning can learn to align the representations from di■

erent views efficiently, especially in cases where the data is not isotropic. In this work, we analyze the training dynamics of a simple multimodal contrastive learning model, and show that contrastive pairs are important for the model to efficiently balance the learned representations. In particular, we reveal a stage-wise behavior of the learning process: In the first stage, the model aligns the feature representations using positive pairs and the condition number grows in this stage. Then, in the second stage, the model reduces the condition number of the learned representations using negative pairs.

****

MAD for Robust Reinforcement Learning in Machine Translation

Domenic Donato,Lei Yu,Wang Ling,Chris Dyer

We introduce a new distributed policy gradient algorithm and show that it outperforms existing reward-aware training procedures such as REINFORCE, minimum risk training (MRT) and proximal policy optimization (PPO) in terms of training stability and generalization performance when optimizing machine translation models. Our algorithm, which we call MAD (on account of using the mean absolute deviation in the importance weighting calculation), has distributed data generators sampling multiple candidates per source sentence on worker nodes, while a central learner updates the policy. MAD depends crucially on two variance reduction strategies: (1) a conditional reward normalization method that ensures each source sentence has both positive and negative reward translation examples and (2) a new robust importance weighting scheme that acts as a conditional entropy regularizer. Experiments on a variety of translation tasks show that policies learned using the MAD algorithm perform very well when using both greedy decoding and beam search, and that the learned policies are sensitive to the specific reward used during training.

****

An Exploration of Conditioning Methods in Graph Neural Networks

Yeskendir Koishekenov,Erik J Bekkers

The flexibility and effectiveness of message passing based graph neural networks (GNNs) induced considerable advances in deep learning on graph-structured data. In such approaches, GNNs recursively update node representations based on their neighbors and they gain expressivity through the use of node and edge attribute vectors. E.g., In computational tasks such as physics and chemistry usage of edge attributes such as relative position or distance proved to be essential. In this work, we address not what kind of attributes to use, but how to condition on this information to improve model performance. We consider three types of conditioning; weak, strong, and pure, which respectively relate to concatenation-based conditioning, gating, and transformations that are causally dependent on the attributes. This categorization provides a unifying viewpoint on different classes of GNNs, from separable convolutions to various forms of message passing networks. We provide an empirical study on the effect of conditioning methods in several tasks in computational chemistry.

****

Speed Up Iterative Non-Autoregressive Transformers by Distilling Multiple Steps

Sajad Norouzi,Rasa Hosseinzadeh,Felipe Perez,Maksims Volkovs

The computational benefits of iterative non-autoregressive transformers decrease as the number of decoding steps increases. As a remedy, we introduce Distill Multiple Steps (DiMS), a simple yet effective distillation technique to decrease the number of required steps to reach a certain translation quality. The distilled model enjoys the computational benefits of early iterations while preserving the enhancements from several iterative steps. DiMS relies on two models namely student and teacher. The student is optimized to predict the output of the teacher after multiple decoding steps while the teacher follows the student via a slow-moving average. The moving average keeps the teacher's knowledge updated and enhances the quality of the labels provided by the teacher. During inference, the student is used for translation and no additional computation is added. We verify the effectiveness of DiMS on various models obtaining 7 and 12.9 BLEU points improvements on distilled and raw versions of WMT'14 De-En, respectively.

****

Cross-Silo Training of Differentially Private Models with Secure Multiparty Computation
Sikha Pentyala,Davis Railsback,Ricardo José Menezes Maia,David Melanson,Rafael Dowsley,Anderson Nascimento,Martine De Cock
We address the problem of learning a machine learning model from training data that originates at multiple data holders in a cross-silo federated setup, while providing formal privacy guarantees regarding the protection of each holder's data. Existing solutions based on Differential Privacy (DP) achieve this at the cost of a drop in accuracy. Solutions based on Secure Multiparty Computation (MPC) do not incur such accuracy loss but leak information when the trained model is made publicly available. We propose an MPC solution for training differentially private models. Our solution relies on an MPC protocol for model training, and an MPC protocol for perturbing the trained model coefficients with Laplace noise in a privacy-preserving manner. The resulting MPC+DP approach achieves higher accuracy than a pure DP approach, while providing the same formal privacy guarantees.
****

HyperTime: Implicit Neural Representations for Time Series Generation
Elizabeth Fons,Alejandro Sztrajman,Yousef El-Laham,Alexandros Iosifidis,Svitlana Vyetrenko
Implicit neural representations (INRs) have recently emerged as a powerful tool that provides an accurate and resolution-independent encoding of data. Their robustness as general approximators has been shown in a wide variety of data sources, with applications on image, sound, and 3D scene representation. However, little attention has been given to leveraging these architectures for the representation and analysis of time series data. In this paper, we propose a new INR architecture for time series (iSIREN) designed to perform an accurate reconstruction of univariate and multivariate data, while also providing an interpretable encoding of the signal. We compare our architecture against SIREN and INRs with different activations, in terms of training convergence, and the reconstruction accuracy of both the signal and its spectral distribution.
To achieve generalization, we propose a hypernetwork architecture (HyperTime) that leverages iSIRENs to learn a latent representation of an entire time series dataset. In addition to the traditional reconstruction loss, we introduce an FFT-based loss that guides the training by enforcing a good match of the ground truth spectral distribution. We show how these architectures can be used for time series generation, and evaluate our method through fidelity metrics, presenting results that exceed the performance of state-of-the-art techniques. Finally, we propose an alternative hypernetwork architecture (iHyperTime) that incorporates interpretability into the latent representation, enabling the introduction of prior knowledge by imposing constraints into the generation process.
****

Homotopy Learning of Parametric Solutions to Constrained Optimization Problems
Shimiao Li,Jan Drgona,Aaron R Tuor,Larry Pileggi,Draguna L Vrabie
Building deep learning (DL) alternatives to constrained optimization problems has been proposed as a cheaper solution approach than classical constrained optimization solvers. However, these approximate learning-based solutions still suffer from constraint violations. From this perspective, reaching a reliable convergence remains an open challenge to DL models even with state-of-the-art methods to impose constraints, especially when facing a large set of nonlinear constraints forming a non-convex feasible set. In this paper, we propose the use of homotopy meta-optimization heuristics which creates a continuous transformation of the objective and constraints during training, to promote a more reliable convergence where the solution feasibility can be further improved. The method developed in this work includes 1) general-purpose homotopy heuristics based on the relaxation of objectives and constraint bounds to enlarge the basin of attraction and 2) physics-informed transformation of domain problem leading to trivial starting points lying within the basin of attraction. Experimentally, we demonstrate the efficacy of the proposed method on a set of abstract constrained optimization problems and real-world power grid optimal power flow problems with increasing co

mplexity. Results show that constrained deep learning models with homotopy heuristics can improve the feasibility of the resulting solutions while achieving near-optimal objective values when compared with non-homotopy counterparts.

****

When Rigid Coherency Hurts: Distributional Coherency Regularization for Probabilistic Hierarchical Time Series Forecasting

Harshavardhan Kamarthi,Lingkai Kong,Alexander Rodríguez,Chao Zhang,B. Aditya Prakash

Probabilistic hierarchical time-series forecasting is an important variant of time-series forecasting, where the goal is to model and forecast multivariate time-series that have hierarchical relations. Previous works all assume rigid consistency over the given hierarchies and do not adapt to real-world data that show deviation from this assumption. Moreover, recent state-of-art neural probabilistic methods also impose hierarchical relations on point predictions and samples of distribution. This does not account for full forecast distributions being coherent with the hierarchy and leads to poorly calibrated forecasts. We close both these gaps and propose PROFHIT, a probabilistic hierarchical forecasting model that jointly models forecast distributions over the entire hierarchy. PROFHIT (1) uses a flexible probabilistic Bayesian approach and (2) introduces soft distributional coherency regularization that enables end-to-end learning of the entire forecast distribution leveraging information from the underlying hierarchy. This enables robust and calibrated forecasts as well as adaptation to real-life data with varied hierarchical consistency. PROFHIT provides 41-88% better performance in accuracy and 23-33% better calibration over a wide range of dataset consistency. Furthermore, PROFHIT can robustly provide reliable forecasts even if up to 10% of input time-series data is missing, whereas other methods’ performance severely degrade by over 70%.

****

MALIBO: Meta-Learning for Likelihood-free Bayesian Optimization

Jiarong Pan,Stefan Falkner,Felix Berkenkamp,Joaquin Vanschoren

Bayesian Optimization (BO) is a popular method to optimize expensive black-box functions. Typically, BO only uses observations from the current task. Recently proposed methods try to warm-start BO by exploiting knowledge from related tasks, yet suffer from scalability issues and sensitivity to heterogeneous scale across multiple datasets. We propose a novel approach to solve these problems by combining a meta-learning technique and a likelihood-free acquisition function. The meta-learning model simultaneously learns the underlying (task-agnostic) data distribution and a latent feature representation for individual tasks. The likelihood-free BO technique has less stringent assumptions about the problems and works with any classification algorithm, making it computation efficient and robust to different scales across tasks. Finally, gradient boosting is used as a residual model on top to adapt to distribution drifts between new and prior tasks, which might otherwise weaken the usefulness of the meta-learned features. Experiments show that the meta-model learns an effective prior for warm-starting optimization algorithms, while being cheap to evaluate and invariant to changes of scale across different datasets.

****

Finding and only finding local Nash equilibria by both pretending to be a follower

Xuchan Bao,Guodong Zhang

Finding (local) Nash equilibria in two-player differentiable games is a classical problem in game theory with important relevance in machine learning. We propose double Follow-the-Ridge (double-FTR), an algorithm that locally converges to and only to local Nash equilibria in general-sum two-player differentiable games. To our knowledge, double-FTR is the first algorithm with such guarantees for general-sum games. Furthermore, we show that by varying its preconditioner, double-FTR leads to a broader family of algorithms with the same convergence guarantee. In addition, double-FTR avoids oscillation near equilibria due to the real-eigenvalues of its Jacobian at fixed points.

Empirically, we validate the double-FTR algorithm on a range of simple zero-sum and general sum games, as well as simple Generative Adversarial Network (GAN) tasks.

****

Learning Low Dimensional State Spaces with Overparameterized Recurrent Neural Nets
Edo Cohen-Karlik,Itamar Menuhin-Gruman,Raja Giryes,Nadav Cohen,Amir Globerson
Overparameterization in deep learning refers to settings where a trained Neural Network (NN) has representational capacity to fit the training data in many ways, some of which generalize well, while others do not. In the case of Recurrent Neural Networks (RNNs) there exists an additional layer of overparameterization, in the sense that a model may exhibit many solutions that generalize well for sequence lengths seen in training, some of which \emph{extrapolate} to longer sequences, while others do not. Numerous works studied the tendency of Gradient Descent (GD) to fit overparameterized NNs with solutions that generalize well. On the other hand, its tendency to fit overparameterized RNNs with solutions that extrapolate has been discovered only lately, and is far less understood. In this paper, we analyze the extrapolation properties of GD when applied to overparameterized linear RNNs. In contrast to recent arguments suggesting an implicit bias towards short-term memory, we provide theoretical evidence for learning low dimensional state spaces, which can also model long-term memory. Our result relies on a dynamical characterization showing that GD (with small step size and near zero initialization) strives to maintain a certain form of balancedness, as well as tools developed in the context of the \emph{moment problem} from statistics (recovery of discrete probability distribution from its moments). Experiments corroborate our theory, demonstrating extrapolation via learning low dimensional state spaces with both linear and non-linear RNNs.

****

Images as Weight Matrices: Sequential Image Generation Through Synaptic Learning Rules
Kazuki Irie,Jürgen Schmidhuber
Work on fast weight programmers has demonstrated the effectiveness of key/value outer product-based learning rules for sequentially generating a weight matrix (WM) of a neural net (NN) by another NN or itself. However, the weight generation steps are typically not visually interpretable by humans, because the contents stored in the WM of an NN are not. Here we apply the same principle to generate natural images. The resulting fast weight painters (FPAs) learn to execute sequences of delta learning rules to sequentially generate images as sums of outer products of self-invented keys and values, one rank at a time, as if each image was a WM of an NN. We train our FPAs in the generative adversarial networks framework, and evaluate on various image datasets. We show how these generic learning rules can generate images with respectable visual quality without any explicit inductive bias for images. While the performance largely lags behind the one of specialised state-of-the-art image generators, our approach allows for visualising how synaptic learning rules iteratively produce complex connection patterns, yielding human-interpretable meaningful images. Finally, we also show that an additional convolutional U-Net (now popular in diffusion models) at the output of an FPA can learn one-step "denoising" of FPA-generated images to enhance their quality.
Our code is public.

****

SurCo: Learning Linear Surrogates for Combinatorial Nonlinear Optimization Problems
Aaron M Ferber,Taoan Huang,Daochen Zha,Martin Schubert,Benoit Steiner,Bistra Dilkina,Yuandong Tian
Optimization problems with expensive nonlinear cost functions and combinatorial constraints appear in many real-world applications, but remain challenging to solve efficiently. Existing combinatorial solvers like Mixed Integer Linear Programming can be fast in practice but cannot readily optimize nonlinear cost functions, while general nonlinear optimizers like gradient descent often do not handle complex combinatorial structures, may require many queries of the cost function

, and are prone to local optima. To bridge this gap, we propose SurCo that learns linear Surrogate costs which can be used by existing Combinatorial solvers to output good solutions to the original nonlinear combinatorial optimization problem, combining the flexibility of gradient-based methods with the structure of linear combinatorial optimization. We learn these linear surrogates end-to-end with the nonlinear loss by differentiating through the linear surrogate solver. Three variants of SurCo are proposed: SurCo-zero operates on individual nonlinear problems, SurCo-prior trains a linear surrogate predictor on distributions of problems, and SurCo-hybrid uses a model trained offline to warm start online solving for SurCo-zero. We analyze our method theoretically and empirically, showing smooth convergence and improved performance. Experiments show that compared to state-of-the-art approaches and expert-designed heuristics, SurCo obtains lower cost solutions with comparable or faster solve time for two real-world industry-level applications: embedding table sharding and inverse photonic design.
****

DT+GNN: A Fully Explainable Graph Neural Network using Decision Trees
Peter Müller,Lukas Faber,Karolis Martinkus,Roger Wattenhofer
We propose a new Decision Tree Graph Neural Network (DT+GNN) architecture for Graph Neural Network (GNN) explanation. Existing post-hoc explanation methods highlight important inputs but fail to reveal how a GNN uses these inputs. In contrast DT+GNN is fully explainable: Humans can inspect and understand the decision making of DT+GNN at every step. DT+GNN internally uses a novel GNN layer that is restricted to categorical state spaces for nodes and messages. After training with gradient descent we can easily distill these layers into decision trees. These trees are further pruned using our newly proposed method to ensure they are small and easy to interpret. DT+GNN can also compute node-level importance scores like the existing explanation methods. We demonstrate on real-world GNN benchmarks that DT+GNN has competitive classification accuracy and computes competitive explanations. Furthermore, we leverage DT+GNN's full explainability to inspect the decision processes in synthetic and real-world datasets with surprising results. We make this inspection accessible through an interactive web tool.
****

Why (and When) does Local SGD Generalize Better than SGD?
Xinran Gu,Kaifeng Lyu,Longbo Huang,Sanjeev Arora
Local SGD is a communication-efficient variant of SGD for large-scale training, where multiple GPUs perform SGD independently and average the model parameters periodically. It has been recently observed that Local SGD can not only achieve the design goal of reducing the communication overhead but also lead to higher test accuracy than the corresponding SGD baseline (Lin et al., 2020b), though the training regimes for this to happen are still in debate (Ortiz et al., 2021). This paper aims to understand why (and when) Local SGD generalizes better based on Stochastic Differential Equation (SDE) approximation. The main contributions of this paper include (i) the derivation of an SDE that captures the long-term behavior of Local SGD in the small learning rate regime, showing how noise drives the iterate to drift and diffuse after it has reached close to the manifold of local minima, (ii) a comparison between the SDEs of Local SGD and SGD, showing that Local SGD induces a stronger drift term that can result in a stronger effect of regularization, e.g., a faster reduction of sharpness, and (iii) empirical evidence validating that having a small learning rate and long enough training time enables the generalization improvement over SGD but removing either of the two conditions leads to no improvement.
****

Function-space regularized Rényi divergences
Jeremiah Birrell,Yannis Pantazis,Paul Dupuis,Luc Rey-Bellet,Markos Katsoulakis
We propose a new family of regularized Rényi divergences parametrized not only by the order α but also by a variational function space. These new objects are defined by taking the infimal convolution of the standard Rényi divergence with the integral probability metric (IPM) associated with the chosen function space. We derive a novel dual variational representation that can be used to construct numerically tractable divergence estimators. This representation avoids ri

sk-sensitive terms and therefore exhibits lower variance, making it well-behaved when $\alpha>1$; this addresses a notable weakness of prior approaches. We prove several properties of these new divergences, showing that they interpolate between the classical Rényi divergences and IPMs. We also study the $\alpha\to\infty$ limit, which leads to a regularized worst-case-regret and a new variational representation in the classical case. Moreover, we show that the proposed regularized Rényi divergences inherit features from IPMs such as the ability to compare distributions that are not absolutely continuous, e.g., empirical measures and distributions with low-dimensional support. We present numerical results on both synthetic and real datasets, showing the utility of these new divergences in both estimation and GAN training applications; in particular, we demonstrate significantly reduced variance and improved training performance.

****

Constant-Factor Approximation Algorithms for Socially Fair k-Clustering

Mehrdad Ghadiri,Mohit Singh,Santosh Vempala

We study approximation algorithms for the socially fair (ℓ_p, k)-clustering problem with m groups which include the socially fair k-median ($p=1$) and k-means ($p=2$). We present (1) a polynomial-time $(5+2\sqrt{6})^p$-approximation with at most $k+m$ centers (2) a $(5+2\sqrt{6}+\epsilon)^p$-approximation with k centers in time $(nk)^{{2^{O(p)} m^2}/\epsilon}$, and (3) a $(15+6\sqrt{6})^p$ approximation with k centers in time $k^{m}\cdot\text{poly}(n)$. The former is obtained by a refinement of the iterative rounding method via a sequence of linear programs. The latter two are obtained by converting a solution with up to $k+m$ centers to one with k centers by sparsification methods for (2) and via an exhaustive search for (3). We also compare the performance of our algorithms with existing approximation algorithms on benchmark datasets, and find that our algorithms outperform existing methods.

****

Implicit Bias of Large Depth Networks: a Notion of Rank for Nonlinear Functions

Arthur Jacot

We show that the representation cost of fully connected neural networks with homogeneous nonlinearities - which describes the implicit bias in function space of networks with L_2-regularization or with losses such as the cross-entropy - converges as the depth of the network goes to infinity to a notion of rank over nonlinear functions. We then inquire under which conditions the global minima of the loss recover the `true' rank of the data: we show that for too large depths the global minimum will be approximately rank 1 (underestimating the rank); we then argue that there is a range of depths which grows with the number of datapoints where the true rank is recovered. Finally, we discuss the effect of the rank of a classifier on the topology of the resulting class boundaries and show that autoencoders with optimal nonlinear rank are naturally denoising.

****

Depth Separation with Multilayer Mean-Field Networks

Yunwei Ren,Mo Zhou,Rong Ge

Depth separation—why a deeper network is more powerful than a shallow one—has been a major problem in deep learning theory. Previous results often focus on representation power, for example, Safran et al. (2019) constructed a function that is easy to approximate using a 3-layer network but not approximable by any 2-layer network. In this paper, we show that this separation is in fact algorithmic: one can learn the function constructed by Safran et al. (2019) using an overparametrized network with polynomially many neurons efficiently. Our result relies on a new way of extending the mean-field limit to multilayer networks, and a decomposition of loss that factors out the error introduced by the discretization of infinite-width mean-field networks.

****

Robust Policy Optimization in Deep Reinforcement Learning

Md Masudur Rahman,Yexiang Xue

Entropy can play an essential role in policy optimization by selecting the stochastic policy, which eventually helps better explore the environment in reinforcement learning (RL). A proper balance between exploration and exploitation is cha llenging and might depend on the particular RL task. However, the stochasticity often reduces as the training progresses; thus, the policy becomes less exploratory. Therefore, in many cases, the policy can converge to sub-optimal due to a lack of representative data during training. Moreover, this issue can even be severe in high-dimensional environments. This paper investigates whether keeping a certain entropy threshold throughout training can help better policy learning. In particular, we propose an algorithm Robust Policy Optimization (RPO), which leverages a perturbed Gaussian distribution to encourage high-entropy actions. We evaluated our methods on various continuous control tasks from DeepMind Control, OpenAI Gym, Pybullet, and IsaacGym. We observed that in many settings, RPO increases the policy entropy early in training and then maintains a certain level of entropy throughout the training period. Eventually, our agent RPO shows consistently improved performance compared to PPO and other techniques such as data augmentation and entropy regularization. Furthermore, in several settings, our method stays robust in performance, while other baseline mechanisms fail to improve and even worsen the performance.

****

Analogical Networks for Memory-Modulated 3D Parsing

Nikolaos Gkanatsios,Mayank Singh,Zhaoyuan Fang,Shubham Tulsiani,Katerina Fragkiadaki

Despite recent breakthroughs in the applications of deep neural networks in visual perception, one setting that presents a persistent challenge is that of “few-shot learning.” Works in the area of few shot visual learning mostly address the task of coarse image classification. Fine-grain visual parsing is necessary for scene understanding and action recognition. Thus far, a separate neural model is trained to parse each semantic category, which hinders knowledge sharing across objects, let alone few shot visual parsing. We present Analogical Networks, a model that casts fine-grained visual parsing into analogical inference: instead of mapping input scenes to part labels, which is hard to adapt in a few-shot manner to novel inputs, our model retrieves related scenes from memory and their corresponding part structures, and predicts analogous part structures in the input scene, via an end-to-end learnable modulation mechanism. By conditioning on more than one memory, compositions of structures are predicted, that mix and match parts from different visual experiences. We show Analogical Networks excel at few-shot learning, where instances of novel object categories are successfully parsed simply by expanding the model’s memory, without any weight updates. Analogical Networks outperform existing state-of-the-art detection transformer models at part segmentation, as well as paradigms of meta-learning and few-shot learning. We show part correspondences emerge across memory and input scenes by simply training for a label-free segmentation objective, as a byproduct of the analogical inductive bias.

****

Fake It Until You Make It : Towards Accurate Near-Distribution Novelty Detection

Hossein Mirzaei,Mohammadreza Salehi,Sajjad Shahabi,Efstratios Gavves,Cees G. M. Snoek,Mohammad Sabokrou,Mohammad Hossein Rohban

We aim for image-based novelty detection. Despite considerable progress, existing models either fail or face dramatic drop under the so-called ``near-distribution" setup, where the differences between normal and anomalous samples are subtle. We first demonstrate existing methods could experience up to 20\% decrease in their AUCs in the near-distribution setting. Next, we propose to exploit a score-based generative model to produce synthetic near-distribution anomalous data. Our model is then fine-tuned to distinguish such data from the normal samples. We make quantitative as well as qualitative evaluation of this strategy, and compare the results with a variety of GAN-based models. Effectiveness of our method for both near-distribution and standard novelty detection is assessed through extensive experiments on datasets in diverse applications such as medical images, object classification, and quality control. This reveals that our method significantly improves upon existing models, and consistently decreases the gap between the near-distribution and standard novelty detection AUCs by a considerable amount.

****

DySR: Adaptive Super-Resolution via Algorithm and System Co-design

Syed Zawad,Cheng Li,Zhewei Yao,Elton Zheng,Yuxiong He,Feng Yan

Super resolution (SR) is a promising approach for improving the quality of low resolution steaming services on mobile devices.

On mobile devices, the available computing and memory resources change dynamically depending on other running applications.

Due to the high computation and memory demands of SR models, it is essential to adapt the model according to available resources to harvest the best possible model performance while maintaining quality of service (QoS), such as meeting a minimum framerate and avoiding interruptions. Nevertheless, there is no SR model or machine learning system that supports adaptive SR, and enabling adaptive SR model on mobile devices is challenging because adapting model can cause significant framerate drop or even service interruption. To address this challenge, we take an algorithm and system co-design approach and propose DySR that maintains QoS while maximizing the model performance. During the training stage, DySR employs an adaption-aware one-shot Neural Architecture Search to produce sub-graphs that share kernel operation weights for low model adaption overhead while striking a balance between performance and framerate. During the inference stage, an incremental model adaption method is developed for further reducing the model adaption overhead. We evaluate on a diverse set of hardware and datasets to show that DySR can generate models close to the Pareto frontier while maintaining a steady framerate throughput with a memory footprint of around 40\% less compared to baseline methods.

****

Domain Invariant Q-Learning for model-free robust continuous control under visual distractions

Tom Dupuis,Jaonary Rabarisoa,Quoc Cuong PHAM,David Filliat

End-to-end reinforcement learning on images showed significant performance progress in the recent years, especially with regularization to value estimation brought by data augmentation \citep{yarats2020image}. At the same time, domain randomization and representation learning helped push the limits of these algorithms in visually diverse environments, full of distractors and spurious noise, making RL more robust to unrelated visual features. We present DIQL, a method that combines risk invariant regularization and domain randomization to reduce out-of-distribution generalization gap for temporal-difference learning. In this work, we draw a link by framing domain randomization as a richer extension of data augmentation to RL and support its generalized use. Our model-free approach improve baselines performances without the need of additional representation learning objectives and with limited additional computational cost. We show that DIQL outperforms existing methods on complex visuo-motor control environment with high visual perturbation. In particular, our approach achieves state-of the-art performance on the Distracting Control Suite benchmark, where we evaluate the robustness to a number of visual perturbators, as well as OOD generalization and extrapolation capabilities.

****

Continual Learning with Soft-Masking of Parameter-Level Gradient Flow

Tatsuya Konishi,Mori Kurokawa,Chihiro Ono,Zixuan Ke,Gyuhak Kim,Bing Liu

Existing research on task incremental learning in continual learning has primarily focused on preventing catastrophic forgetting (CF). Several techniques have achieved learning with no CF. However, they attain it by letting each task monopolize a sub-network in a shared network, which seriously limits knowledge transfer (KT) and causes over-consumption of the network capacity, i.e., as more tasks are learned, the performance deteriorates. The goal of this paper is threefold: (1) overcoming CF, (2) encouraging KT, and (3) tackling the capacity problem. A novel and simple technique (called SPG) is proposed that soft-masks (partially blocks) parameter updating in training based on the importance of each parameter to old tasks. Each task still uses the full network, i.e., no monopoly of any part of the network by any task, which enables maximum KT and reduction of capacity usage. Extensive experiments demonstrate the effectiveness of SPG in achieving

all three objectives. More notably, it attains significant transfer of knowledge not only among similar tasks (with shared knowledge) but also among dissimilar tasks (with little shared knowledge) while preventing CF.
****

Asynchronous Message Passing: A new Framework for Learning in Graphs
Lukas Faber,Roger Wattenhofer
This paper studies asynchronous message passing (AMP), a new framework for applying neural networks to graphs. Existing graph neural networks (GNNs) use the message passing framework which is based on the synchronous distributed computing model. In traditional GNNs, nodes aggregate their neighbors in each round, which causes problems such as oversmoothing and expressiveness limitations. On the other hand, our AMP framework is based on the \textit{asynchronous} model, where nodes react to messages of their neighbors individually. We prove (i) AMP is at least as powerful as the message passing framework, (ii) AMP is more powerful than the $1-$WL test for graph isomorphism, an important benchmark for message passing GNNs, and (iii) conceptually, AMP can even separate any pair of graphs and compute graph isomorphism. We experimentally validate the findings on AMP's expressiveness, and show that AMP might be better suited to propagate messages over large distances in graphs. We also demonstrate that AMP performs well on several graph classification benchmarks.
****

Integrating Symmetry into Differentiable Planning with Steerable Convolutions
Linfeng Zhao,Xupeng Zhu,Lingzhi Kong,Robin Walters,Lawson L.S. Wong
To achieve this, we draw inspiration from equivariant convolution networks and model the path planning problem as a set of signals over grids. We demonstrate that value iteration can be treated as a linear equivariant operator, which is effectively a steerable convolution. Building upon Value Iteration Networks (VIN), we propose a new Symmetric Planning (SymPlan) framework that incorporates rotation and reflection symmetry using steerable convolution networks. We evaluate our approach on four tasks: 2D navigation, visual navigation, 2 degrees of freedom (2-DOF) configuration space manipulation, and 2-DOF workspace manipulation. Our experimental results show that our symmetric planning algorithms significantly improve training efficiency and generalization performance compared to non-equivariant baselines, including VINs and GPPN.
****

MolJET: Multimodal Joint Embedding Transformer for Conditional de novo Molecular Design and Multi-Property Optimization
Orion Walker Dollar,Sameera Horawalavithana,Scott Vasquez,W. James Pfaendtner,Svitlana Volkova
Multi-property constrained optimization of molecules using generative de novo design models is vital for the successful application of Artificial Intelligence (AI) towards materials and drug discovery. Yet there remains a gap between the reported performance of such models in the literature and their practical utility in real world design scenarios. Furthermore, existing models are largely inaccessible to chemists without an extensive background in computer science. To address these challenges, we propose a generative foundation model, the Multimodal Joint Embedding Transformer (MolJET), which performs conditional generation of desired molecular distributions based on human-interpretable chemistry prompts in a zero-shot manner. We assess MolJET on the standard benchmarks available in the GuacaMol and MIMOSA evaluation frameworks. These include structure-based sampling tasks as well as a range of multi-property optimization tasks that probe a models ability to design drug-like molecules given realistic property constraints. We demonstrate that with self-supervised pretraining, MolJET outperforms 80% of task-optimized models while using zero-shot inferences and beats all baselines after minimal supervision. Moreover, the performance of MolJET on text-only conditioning tasks improves with the inclusion of property modalities during training, highlighting the importance of a multimodal approach to molecular design. MolJET is the first example of text-based de novo molecular design using large-scale multimodal foundation models and should serve as a building block towards further improvements to accessible AI for chemists.

****

The Challenges of Exploration for Offline Reinforcement Learning

Nathan Lambert,Markus Wulfmeier,William F Whitney,Arunkumar Byravan,Michael Bloesch,Vibhavari Dasagi,Tim Hertweck,Martin Riedmiller

Offline Reinforcement Learning (ORL) enables us to separately study the two interlinked processes of reinforcement learning: collecting informative experience and inferring optimal behaviour. The second step has been widely studied in the offline setting, but just as critical to data-efficient RL is the collection of informative data. The task-agnostic setting for data collection, where the task is not known a priori, is of particular interest due to the possibility of collecting a single dataset and using it to solve several downstream tasks as they arise. We investigate this setting via curiosity-based intrinsic motivation, a family of exploration methods which encourage the agent to explore those states or transitions it has not yet learned to model. With Explore2Offline, we propose to evaluate the quality of collected data by transferring the collected data and inferring policies with reward relabelling and standard offline RL algorithms. We evaluate a wide variety of data collection strategies, including a new exploration agent, Intrinsic Model Predictive Control (IMPC), using this scheme and demonstrate their performance on various tasks. We use this decoupled framework to strengthen intuitions about exploration and the data prerequisites for effective offline RL.

****

SGD with large step sizes learns sparse features

Maksym Andriushchenko,Aditya Vardhan Varre,Loucas Pillaud-Vivien,Nicolas Flammarion

We showcase important features of the dynamics of the Stochastic Gradient Descent (SGD) in the training of neural networks. We present empirical observations that the commonly used large step sizes (i) lead the iterates to jump from one side of a valley to the other causing \textit{loss stabilisation} (ii) this stabilisation induces a hidden stochastic dynamics orthogonal to the bouncing directions that \textit{biases it implicitly} toward simple predictors. Furthermore, we show empirically that the longer large step sizes keep SGD high in the loss landscape valleys, the better the implicit regularization can operate and find sparse representations. Notably, no explicit regularization is used so that the regularization effect comes solely from the SGD training dynamics influenced by the step size schedule. Therefore, these observations unveil how, through the step size schedules, both gradient and noise drive together the SGD dynamics through the loss landscape of neural networks. We justify these findings theoretically through the study of simple neural network models. Finally, we shed a new light on some common practice and observed phenomena when training neural networks.

****

Synergies Between Disentanglement and Sparsity: a Multi-Task Learning Perspective

Sebastien Lachapelle,Tristan Deleu,Divyat Mahajan,Ioannis Mitliagkas,Yoshua Bengio,Simon Lacoste-Julien,Quentin Bertrand

Although disentangled representations are often said to be beneficial for downstream tasks, current empirical and theoretical understanding is limited. In this work, we provide evidence that disentangled representations coupled with sparse base-predictors improve generalization. In the context of multi-task learning, we prove a new identifiability result that provides conditions under which maximally sparse base-predictors yield disentangled representations. Motivated by this theoretical result, we propose a practical approach to learn disentangled representations based on a sparsity-promoting bi-level optimization problem. Finally, we explore a meta-learning version of this algorithm based on group Lasso multiclass SVM base-predictors, for which we derive a tractable dual formulation. It obtains competitive results on standard few-shot classification benchmarks, while each task is using only a fraction of the learned representations.

****

Discerning Hydroclimatic Behavior with a Deep Convolutional Residual Regressive Neural Network

Albert Larson,Ali S Akanda,Abdeltawab Hendawi,Soni Pradhanang,Thomas Boving

Water impacts the globe daily in new and familiar ways such as the ongoing western United States drought and the 2022 Pakistan flood. These events sustain uncertainty, risk, and loss forces to the global ecosystem. Better forecasting tools are mandatory to calibrate our response in an effort to mitigate such natural hazards in our watersheds and adapt to the planet’s dynamic environment. Here, we present a Deep Convolutional Residual Regressive Neural Net (DCRRNN - pronounced “discern”) platform for obtaining, visualizing, and analyzing the basin response of watersheds to water cycle fluxes. We examine four very large basins, simulating river response to the hydroclimatic fluxes they face. Experiments modulating the lever of time lag between remotely sensed and ground truth measurements are performed to assess the metrological limits of this forecasting device. The resultant grand mean Nash Sutcliffe and Kling Gupta efficiency values are both of greater value than 90\%. Our results show that DCRRNN can become a powerful resource to simulate and forecast the impacts of hydroclimatic events as they relate to watershed response in a globally changing climate.

****

Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning

Matthew Ashman,Chao Ma,Agrin Hilmkil,Joel Jennings,Cheng Zhang

Latent confounding has been a long-standing obstacle for causal reasoning from observational data. One popular approach is to model the data using acyclic directed mixed graphs (ADMGs), which describe ancestral relations between variables using directed and bidirected edges. However, existing methods using ADMGs are based on either linear functional assumptions or a discrete search that is complicated to use and lacks computational tractability for large datasets. In this work, we further extend the existing body of work and develop a novel gradient-based approach to learning an ADMG with nonlinear functional relations from observational data. We first show that the presence of latent confounding is identifiable under the assumptions of bow-free ADMGs with nonlinear additive noise models. With this insight, we propose a novel neural causal model based on autoregressive flows. This not only enables us to model complex causal relationships behind the data, but also estimate their functional relationships (hence treatment effects) simultaneously. We further validate our approach via experiments on both synthetic and real-world datasets, and demonstrate the competitive performance against relevant baselines.

****

Mitigating Gradient Bias in Multi-objective Learning: A Provably Convergent Approach

Heshan Devaka Fernando,Han Shen,Miao Liu,Subhajit Chaudhury,Keerthiram Murugesan,Tianyi Chen

Many machine learning problems today have multiple objective functions. They appear either in learning with multiple criteria where learning has to make a trade-off between multiple performance metrics such as fairness, safety and accuracy; or, in multi-task learning where multiple tasks are optimized jointly, sharing inductive bias between them. This problems are often tackled by the multi-objective optimization framework. However, existing stochastic multi-objective gradient methods and its variants (e.g., MGDA, PCGrad, CAGrad, etc.) all adopt a biased noisy gradient direction, which leads to degraded empirical performance.

To this end, we develop a stochastic multi-objective gradient correction (MoCo) method for multi-objective optimization. The unique feature of our method is that it can guarantee convergence without increasing the batch size even in the nonconvex setting. Simulations on multi-task supervised and reinforcement learning demonstrate the effectiveness of our method relative to the state-of-the-art methods.

****

Pareto Rank-Preserving Supernetwork for HW-NAS

Hadjer Benmeziane,Hamza Ouarnoughi,Smail Niar,Kaoutar El Maghraoui

In neural architecture search (NAS), training every sampled architecture is very time-consuming and should be avoided.

Weight-sharing is a promising solution to speed up the evaluation process.

However, a sampled subnetwork is not guaranteed to be estimated precisely unless a complete individual training process is done.

Additionally, practical deep learning engineering processes require incorporating realistic hardware-performance metrics into the NAS evaluation process, also known as hardware-aware NAS (HW-NAS).

HW-NAS results a Pareto front, a set of all architectures that optimize conflicting objectives, i.e. task-specific performance and hardware efficiency.

This paper proposes a supernetwork training methodology that preserves the Pareto ranking between its different subnetworks resulting in more efficient and accurate neural networks for a variety of hardware platforms. The results show a 97% near Pareto front approximation in less than 2 GPU days of search, which provides x2 speed up compared to state-of-the-art methods. We validate our methodology on NAS-Bench-201, DARTS and ImageNet. Our optimal model achieves 77.2% accuracy (+1.7% compared to baseline) with an inference time of 3.68ms on Edge GPU for ImageNet.

****

ProSampler: Improving Contrastive Learning by Better Mini-batch Sampling

Zhen Yang,Tinglin Huang,Ming Ding,Zhitao Ying,Yukuo Cen,Yangliao Geng,Yuxiao Dong,Jie Tang

In-batch contrastive learning has emerged as a state-of-the-art self-supervised learning solution, with the philosophy of bringing semantically similar instances closer while pushing dissimilar instances apart within a mini-batch. However, the in-batch negative sharing strategy is limited by the batch size and falls short of prioritizing the informative negatives (i.e., hard negatives) globally. In this paper, we propose to sample mini-batches with hard negatives on a proximity graph in which the instances (nodes) are connected according to the similarity measurement. Sampling on the proximity graph can better exploit the hard negatives globally by bridging in similar instances from the entire dataset. The proposed method can flexibly explore the negatives by modulating two parameters, and we show that such flexibility is the key to better exploit hard negative globally. We evaluate the proposed method on three representative contrastive learning algorithms, each of which corresponds to one modality: image, text, and graph. Besides, we also apply it to the variants of the InfoNCE objective to verify its generality. The results show that our method can consistently boost the performance of contrastive methods, with a relative improvement of 2.5% for SimCLR on ImageNet-100, 1.4% for SimCSE on the standard STS task, and 1.2% for GraphCL on the COLLAB dataset.

****

$O(T^{-1})$ Convergence of Optimistic-Follow-the-Regularized-Leader in Two-Player Zero-Sum Markov Games

Yuepeng Yang,Cong Ma

We prove that the optimistic-follow-the-regularized-leader (OFTRL) algorithm, together with smooth value updates, finds an $O(T^{-1})$ approximate Nash equilibrium in T iterations for two-player zero-sum Markov games with full information. This improves the $\tilde{O}(T^{-5/6})$ convergence rate recently shown by Zhang et al (2022). The refined analysis hinges on two essential ingredients. First, the sum of the regrets of the two players, though not necessarily non-negative as in normal-form games, is approximately non-negative in Markov games. This property allows us to bound the second-order path lengths of the learning dynamics. Second, we prove a tighter algebraic inequality regarding the weights deployed by OFTRL that shaves an extra $\log T$ factor. This crucial improvement enables the inductive analysis that leads to the final $O(T^{-1})$ rate.

****

Bispectral Neural Networks

Sophia Sanborn,Christian A Shewmake,Bruno Olshausen,Christopher J. Hillar

We present a neural network architecture, Bispectral Neural Networks (BNNs) for learning representations that are invariant to the actions of compact commutative groups on the space over which a signal is defined. The model incorporates the ansatz of the bispectrum, an analytically defined group invariant that is complete -- that is, it preserves all signal structure while removing only the variat ion due to group actions. Here, we demonstrate that BNNs are able to simultaneously learn groups, their irreducible representations, and corresponding equivariant and complete-invariant maps purely from the symmetries implicit in data. Further, we demonstrate that the completeness property endows these networks with strong invariance-based adversarial robustness. This work establishes Bispectral Neural Networks as a powerful computational primitive for robust invariant representation learning.

****

Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise

Arpit Bansal,Eitan Borgnia,Hong-Min Chu,Jie S. Li,Hamid Kazemi,Furong Huang,Micah Goldblum,Jonas Geiping,Tom Goldstein

Standard diffusion models involve an image transform -- adding Gaussian noise -- and an image restoration operator that inverts this degradation. We observe that the generative behavior of diffusion models is not strongly dependent on the choice of image degradation, and in fact an entire family of generative models can be constructed by varying this choice. Even when using completely deterministic degradations (e.g., blur, masking, and more), the training and test-time update rules that underlie diffusion models can be easily generalized to create generative models. The success of these fully deterministic models calls into question the community's understanding of diffusion models, which relies on noise in either gradient Langevin dynamics or variational inference, and paves the way for generalized diffusion models that invert arbitrary processes.

****

Beyond Lipschitz: Sharp Generalization and Excess Risk Bounds for Full-Batch GD

Konstantinos Nikolakakis,Farzin Haddadpour,Amin Karbasi,Dionysios Kalogerias

We provide sharp path-dependent generalization and excess risk guarantees for the full-batch Gradient Descent (GD) algorithm on smooth losses (possibly non-Lipschitz, possibly nonconvex). At the heart of our analysis is an upper bound on the generalization error, which implies that average output stability and a bounded expected optimization error at termination lead to generalization. This result shows that a small generalization error occurs along the optimization path, and allows us to bypass Lipschitz or sub-Gaussian assumptions on the loss prevalent in previous works. For nonconvex, convex, and strongly convex losses, we show the explicit dependence of the generalization error in terms of the accumulated path-dependent optimization error, terminal optimization error, number of samples, and number of iterations. For nonconvex smooth losses, we prove that full-batch GD efficiently generalizes close to any stationary point at termination, and recovers the generalization error guarantees of stochastic algorithms with fewer assumptions. For smooth convex losses, we show that the generalization error is tighter than existing bounds for SGD (up to one order of error magnitude). Consequently the excess risk matches that of SGD for quadratically less iterations. Lastly, for strongly convex smooth losses, we show that full-batch GD achieves essentially the same excess risk rate as compared with the state of the art on SGD, but with an exponentially smaller number of iterations (logarithmic in the dataset size).

****

Zero-Shot Retrieval with Search Agents and Hybrid Environments

Michelle Chen Huebscher,Christian Buck,Massimiliano Ciaramita,Sascha Rothe

Learning to search is the task of building artificial agents that learn to autonomously use a search box to find information. So far, it has been shown that current language models can learn symbolic query reformulation policies, in combination with traditional term-based retrieval, but fall short of outperforming neural retrievers. We extend the previous learning to search setup to a hybrid environment, which accepts discrete query refinement operations, after a first-pass retrieval step performed by a dual encoder. Experiments on the BEIR task show that search agents, trained via behavioral cloning, outperform the underlying search system based on a combined dual encoder retriever and cross encoder reranker. Furthermore, we find that simple heuristic Hybrid Retrieval Environments (HRE) can improve baseline performance by several nDCG points. The search agent based on the HRE environment (HaRE) produces state-of-the-art performance on both zero-shot and in-domain evaluations. We carry out an extensive qualitative analysis to shed light on the agents policies.

****

Hyper-Decision Transformer for Efficient Online Policy Adaptation

Mengdi Xu,Yuchen Lu,Yikang Shen,Shun Zhang,Ding Zhao,Chuang Gan

Decision Transformers (DT) have demonstrated strong performances in offline reinforcement learning settings, but quickly adapting to unseen novel tasks remains challenging. To address this challenge, we propose a new framework, called Hyper-Decision Transformer (HDT), that can generalize to novel tasks from a handful of demonstrations in a data- and parameter-efficient manner. To achieve such a goal, we propose to augment the base DT with an adaptation module, whose parameters are initialized by a hyper-network. When encountering unseen tasks, the hyper-network takes a handful of demonstrations as inputs and initializes the adaptation module accordingly. This initialization enables HDT to efficiently adapt to novel tasks by only fine-tuning the adaptation module. We validate HDT's generalization capability on object manipulation tasks. We find that with a single expert demonstration and fine-tuning only 0.5% of DT parameters, HDT adapts faster to unseen tasks than fine-tuning the whole DT model. Finally, we explore a more challenging setting where expert actions are not available, and we show that HDT outperforms state-of-the-art baselines in terms of task success rates by a large margin. Demos are available on our project page: https://sites.google.com/view/hdtforiclr2023/home.

****

Deep Learning of Intrinsically Motivated Options in the Arcade Learning Environment

Louis Bagot,Kevin Mets,Tom De Schepper,Steven Latre

In Reinforcement Learning, Intrinsic Motivation motivates directed behaviors through a wide range of reward-generating methods. Depending on the task and environment, these rewards can be useful, might complement each other, but can also break down entirely, as seen with the noisy TV problem for curiosity. We therefore argue that scalability and robustness, among others, are key desirable properties of a method to incorporate intrinsic rewards, which a simple weighted sum of reward lacks. In a tabular setting, Explore Options let the agent call an intrinsically motivated policy in order to learn from its trajectories. We introduce Deep Explore Options, revising Explore Options within the Deep Reinforcement Learning paradigm to tackle complex visual problems. Deep Explore Options can naturally learn from several unrelated intrinsic rewards, ignore harmful intrinsic rewards, learn to balance exploration, but also isolate exploitative and exploratory behaviors for independent usage.

We test Deep Explore Options on hard and easy exploration games of the Atari Suite, following a benchmarking study to ensure fairness. Our empirical results show that they achieve similar results than weighted sum baselines, while maintaining their key properties.

****

Solving Continuous Control via Q-learning

Tim Seyde,Peter Werner,Wilko Schwarting,Igor Gilitschenski,Martin Riedmiller,Daniela Rus,Markus Wulfmeier

While there has been substantial success for solving continuous control with actor-critic methods, simpler critic-only methods such as Q-learning find limited application in the associated high-dimensional action spaces. However, most actor-critic methods come at the cost of added complexity: heuristics for stabilisation, compute requirements and wider hyperparameter search spaces. We show that a simple modification of deep Q-learning largely alleviates these issues. By combining bang-bang action discretization with value decomposition, framing single-agent control as cooperative multi-agent reinforcement learning (MARL), this simple critic-only approach matches performance of state-of-the-art continuous actor-critic methods when learning from features or pixels. We extend classical bandit examples from cooperative MARL to provide intuition for how decoupled critics leverage state information to coordinate joint optimization, and demonstrate surp risingly strong performance across a variety of continuous control tasks.

****

Make-A-Video: Text-to-Video Generation without Text-Video Data

Uriel Singer,Adam Polyak,Thomas Hayes,Xi Yin,Jie An,Songyang Zhang,Qiyuan Hu,Harry Yang,Oron Ashual,Oran Gafni,Devi Parikh,Sonal Gupta,Yaniv Taigman

We propose Make-A-Video -- an approach for directly translating the tremendous recent progress in Text-to-Image (T2I) generation to Text-to-Video (T2V). Our intuition is simple: learn what the world looks like and how it is described from paired text-image data, and learn how the world moves from unsupervised video footage. Make-A-Video has three advantages: (1) it accelerates training of the T2V model (it does not need to learn visual and multimodal representations from scratch), (2) it does not require paired text-video data, and (3) the generated videos inherit the vastness (diversity in aesthetic, fantastical depictions, etc.) of today's image generation models.

We design a simple yet effective way to build on T2I models with novel and effective spatial-temporal modules. First, we decompose the full temporal U-Net and attention tensors and approximate them in space and time. Second, we design a spatial temporal pipeline to generate high resolution and frame rate videos with a video decoder, interpolation model and two super resolution models that can enable various applications besides T2V. In all aspects, spatial and temporal resolution, faithfulness to text, and quality, Make-A-Video sets the new state-of-the-art in text-to-video generation, as determined by both qualitative and quantitative measures.

****

Unsupervised Adaptation for Fairness under Covariate Shift

Jatin Chauhan,Shreyas Havaldar,Karthikeyan Shanmugam,Jay Nandy,Aravindan Raghuveer

Training fair models typically involves optimizing a composite objective accounting for both prediction accuracy and some fairness measure. However, due to a shift in the distribution of the covariates at test time, the learnt fairness tradeoffs may no longer be valid, which we verify experimentally. To address this, we consider an unsupervised adaptation problem of training fair classifiers when only a small set of unlabeled test samples is available along with a large labeled training set. We propose a novel modification to the traditional composite objective by adding a weighted entropy objective on the unlabeled test dataset. This involves a min-max optimization where weights are optimized to mimic the importance weighting ratios followed by classifier optimization. We demonstrate that our weighted entropy objective provides an upper bound on the standard importance sampled training objective common in covariate shift formulations under some mild conditions. Experimentally, we demonstrate that Wasserstein distance based penalty for representation matching across protected sub groups together with the above loss outperforms existing baselines. Our method achieves the best accuracy-equalized odds tradeoff under the covariate shift setup. We find that, for the same accuracy, we get upto 2x improvement in equalized odds on notable benchmarks.

****

Pushing the limits of self-supervised learning: Can we outperform supervised learning without labels?

Nenad Tomasev,Ioana Bica,Brian McWilliams,Lars Holger Buesing,Razvan Pascanu,Charles Blundell,Jovana Mitrovic

Despite recent progress made by self-supervised methods in representation learning with residual networks, they still underperform supervised learning on the ImageNet classification benchmark, limiting their applicability in performance critical settings. Building on prior theoretical insights from RELIC [Mitrovic et al., 2021], we include additional inductive biases into self-supervised learning. We propose a new self-supervised representation learning method, RELICv2,which combines an explicit invariance loss with a contrastive objective over avaried set of appropriately constructed data views to avoid learning spurious cor-relations and obtain more informative representations. RELICv2 achieves 77.1% top-1 classification accuracy on ImageNet using linear evalu ation with a ResNet50 architecture and 80.6% with larger ResNet models, outperforming previous state-of-the-art self-supervised approaches by a wide margin. Most notably, RELICv2 is the first unsupervised representation learning method to consistently outperform the supervised baseline in a like-for-like comparison over a range of ResNet architectures. Finally, we show that despite using ResNet encoders, RELICv2 is comparable to state-of-the-art self-supervised vision transformers.

****

Personalized Reward Learning with Interaction-Grounded Learning (IGL)

Jessica Maghakian,Paul Mineiro,Kishan Panaganti,Mark Rucker,Akanksha Saran,Cheng Tan

In an era of countless content offerings, recommender systems alleviate information overload by providing users with personalized content suggestions. Due to the scarcity of explicit user feedback, modern recommender systems typically optimize for the same fixed combination of implicit feedback signals across all users. However, this approach disregards a growing body of work highlighting that (i) implicit signals can be used by users in diverse ways, signaling anything from satisfaction to active dislike, and (ii) different users communicate preferences in different ways. We propose applying the recent Interaction Grounded Learning (IGL) paradigm to address the challenge of learning representations of diverse user communication modalities. Rather than requiring a fixed, human-designed reward function, IGL is able to learn personalized reward functions for different users and then optimize directly for the latent user satisfaction. We demonstrate the success of IGL with experiments using simulations as well as with real-world production traces.

****

From Adaptive Query Release to Machine Unlearning

Enayat Ullah,Raman Arora

We formalize the problem of machine unlearning as design of efficient unlearning algorithms corresponding to learning algorithms which perform a selection of adaptive queries from structured query classes. We give efficient unlearning algorithms for linear and prefix-sum query classes. As applications, we show that unlearning in many problems, in particular, stochastic convex optimization (SCO), can be reduced to the above, yielding improved guarantees for the problem. In particular, for smooth Lipschitz losses and any $\rho>0$, our results yield an unlearning algorithm with excess population risk of $\tilde O\big(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{n\rho}\big)$ with unlearning query (gradient) complexity $\tilde O(\rho \cdot \text{Retraining Complexity})$, where d is the model dimensionality and n is the initial number of samples. For non-smooth Lipschitz losses, we give an unlearning algorithm with excess population risk $\tilde O\big(\frac{1}{\sqrt{n}}+\big(\frac{\sqrt{d}}{n\rho}\big)^{1/2}\big)$ with the same unlearning query (gradient) complexity. Furthermore, in the special case of Generalized Linear Models (GLMs), such as those in linear and logistic regression, we get dimension-independent rates of $\tilde O\big(\frac{1}{\sqrt{n}} +\frac{1}{(n\rho)^{2/3}}\big)$ and $\tilde O\big(\frac{1}{\sqrt{n}} +\frac{1}{(n\rho)^{1/3}}\big)$ for smooth Lipschitz and non-smooth Lipschitz losses respectively. Finally, we give generalizations of the above from one unlearning request to dynamic streams consisting of insertions and deletions.

****

ReAct: Synergizing Reasoning and Acting in Language Models

Shunyu Yao,Jeffrey Zhao,Dian Yu,Nan Du,Izhak Shafran,Karthik R Narasimhan,Yuan Cao

While large language models (LLMs) have demonstrated impressive capabilities across tasks in language understanding and interactive decision making, their abilities for reasoning (e.g. chain-of-thought prompting) and acting (e.g. action plan generation) have primarily been studied as separate topics. In this paper, we explore the use of LLMs to generate both reasoning traces and task-specific actions in an interleaved manner, allowing for greater synergy between the two: reasoning traces help the model induce, track, and update action plans as well as ha ndle exceptions, while actions allow it to interface with external sources, such as knowledge bases or environments, to gather additional information. We apply our approach, named ReAct, to a diverse set of language and decision making tasks and demonstrate its effectiveness over state-of-the-art baselines, as well as improved human interpretability and trustworthiness over methods without reasoning or acting components. Concretely, on question answering (HotpotQA) and fact verification (Fever), ReAct overcomes issues of hallucination and error propagation prevalent in chain-of-thought reasoning by interacting with a simple Wikipedia API, and generates human-like task-solving trajectories that are more interpretable than baselines without reasoning traces. On two interactive decision making benchmarks (ALFWorld and WebShop), ReAct outperforms imitation and reinforcement learning methods by an absolute success rate of 34% and 10% respectively, while being prompted with only one or two in-context examples.

****

Towards convergence to Nash equilibria in two-team zero-sum games
Fivos Kalogiannis,Ioannis Panageas,Emmanouil-Vasileios Vlatakis-Gkaragkounis

Contemporary applications of machine learning raise important and overlooked theoretical questions regarding optimization in two-team games. Formally, two-team zero-sum games are defined as multi-player games where players are split into two competing sets of agents, each experiencing a utility identical to that of their teammates and opposite to that of the opposing team. We focus on the solution concept of Nash equilibria and prove $\textrm{CLS}$-hardness of computing them in this class of games. To further examine the capabilities of online learning algorithms in games with full-information feedback, we propose a benchmark of a simple ---yet nontrivial--- family of such games. These games do not enjoy the properties used to prove convergence for relevant algorithms. In particular, we use a dynamical systems perspective to demonstrate that gradient descent-ascent, its optimistic variant, optimistic multiplicative weights update, and extra gradient fail to converge (even locally) to a Nash equilibrium. On a brighter note, we propose a first-order method that leverages control theory techniques and under some conditions enjoys last-iterate local convergence to a Nash equilibrium. We also believe our proposed method is of independent interest for general min-max optimization.

****

Ensemble Homomorphic Encrypted Data Classification
Dana R Alsagheer,Hadi Mansouifar,lei Xu,Qian Lou,Weidong Shi,lin chen

Homomorphic encryption (HE) is encryption that permits users to perform computations on encrypted data without first decrypting it. HE can be used for privacy-preserving outsourced computation and analysis, allowing data to be encrypted and outsourced to commercial cloud environments for processing while encrypted or sensitive data. HE enables new services by removing privacy barriers inhibiting data sharing or increasing the security of existing services. A convolution neural network (CNN) with shallow architecture can be homomorphically evaluated using addition and multiplication by replacing the activation function, such as ReLU, with a low polynomial degree. To achieve the same performance as the ReLU activation function, we study the impact of applying the ensemble techniques to solve the accuracy problem. Our experimental results empirically show that the ensemble approach can reduce bias, and variance, increasing accuracy to achieve the same ReLU performance with parallel and sequential techniques. We demonstrate the effectiveness and robustness of our method using three data sets: MNIST, FMNIST, and CIFAR-10

****

Generative Pretraining for Black-Box Optimization
Siddarth Krishnamoorthy,Satvik Mehul Mashkaria,Aditya Grover

Many problems in science and engineering involve optimizing an expensive black-box function over a high-dimensional space. For such black-box optimization (BBO) problems, we typically assume a small budget for online function evaluations, but also often have access to a fixed, offline dataset for pretraining. Prior approaches seek to utilize the offline data to approximate the function or its inverse but are not sufficiently accurate far from the data distribution. We propose

BONET, a generative framework for pretraining a novel black-box optimizer using offline datasets. In BONET, we train an autoregressive model on fixed-length trajectories corresponding to runs of implicit black-box function optimizers. We design a sampling strategy to synthesize trajectories from offline data using a simple heuristic of rolling out monotonic transitions from low-fidelity to high-fidelity samples. Empirically, we instantiate BONET using a causally masked Transformer and evaluate it on Design-Bench, where we rank the best on average, outperforming state-of-the-art baselines.
****

Discovering Evolution Strategies via Meta-Black-Box Optimization
Robert Tjarko Lange,Tom Schaul,Yutian Chen,Tom Zahavy,Valentin Dalibard,Chris Lu,Satinder Singh,Sebastian Flennerhag
Optimizing functions without access to gradients is the remit of black-box methods such as evolution strategies. While highly general, their learning dynamics are often times heuristic and inflexible — exactly the limitations that meta-learning can address. Hence, we propose to discover effective update rules for evolution strategies via meta-learning. Concretely, our approach employs a search strategy parametrized by a self-attention-based architecture, which guarantees the update rule is invariant to the ordering of the candidate solutions. We show that meta-evolving this system on a small set of representative low-dimensional analytic optimization problems is sufficient to discover new evolution strategies capable of generalizing to unseen optimization problems, population sizes and optimization horizons. Furthermore, the same learned evolution strategy can outperform established neuroevolution baselines on supervised and continuous control tasks. As additional contributions, we ablate the individual neural network components of our method; reverse engineer the learned strategy into an explicit heuristic form, which remains highly competitive; and show that it is possible to self-referentially train an evolution strategy from scratch, with the learned update rule used to drive the outer meta-learning loop.
****

The Use of Open-Source Boards for Data Collection and Machine Learning in Remote Deployments
Gabriel Kiarie,Jason Kabi,Lorna Mugambi,Ciira wa Maina
Machine learning is being adopted in many walks of life to solve various problems. This is being driven by development of robust machine learning algorithms, availability of large datasets and low cost computation resources. Some machine learning applications require deployment of devices off-the-grid for data collection and real time monitoring. Such applications require development of systems that can operate autonomously during their deployment. Advancement in technology has seen development of low-cost and low-power open-source microcontrollers and single board computers. These boards can be interfaced with a wide array of sensors and can perform computation processes. The boards are finding wide applications in data collection and machine learning initiatives. This paper will describe how the boards are leveraged for off-grid deployments.
****

Rhino: Deep Causal Temporal Relationship Learning with History-dependent Noise
Wenbo Gong,Joel Jennings,Cheng Zhang,Nick Pawlowski
Discovering causal relationships between different variables from time series data has been a long-standing challenge for many domains. For example, in stock markets, the announcement of acquisitions from leading companies may have immediate effects on stock prices and increase the uncertainty of the future market due to this past action. To discover causal relations in such case, the model needs to consider non-linear relations between variables, instantaneous effect and the change of noise distribution due to past actions. We name the latter as history-dependent noise. However, previous works do not offer a solution addressing all these problems together. In this paper, we propose a structural equation model, called Rhino, which combines vector auto-regression, deep learning and variational inference to model non-linear relationships with instantaneous effects while allowing the noise distribution to be modulated by history observations. Theoretically, we prove the structural identifiability of Rhino. Our empirical results

from extensive synthetic experiments and two real-world benchmarks demonstrate better discovery performance compared to relevant baselines, with ablation studies revealing its robustness under model misspecification.

****

DensePure: Understanding Diffusion Models for Adversarial Robustness

Chaowei Xiao,Zhongzhu Chen,Kun Jin,Jiongxiao Wang,Weili Nie,Mingyan Liu,Anima Anandkumar,Bo Li,Dawn Song

Diffusion models have been recently employed to improve certified robustness through the process of denoising. However, the theoretical understanding of why diffusion models are able to improve the certified robustness is still lacking, preventing from further improvement. In this study, we close this gap by analyzing the fundamental properties of diffusion models and establishing the conditions under which they can enhance certified robustness. This deeper understanding allows us to propose a new method DensePure, designed to improve the certified robustness of a pretrained model (i.e. classifier). Given an (adversarial) input, DensePure consists of multiple runs of denoising via the reverse process of the diffusion model (with different random seeds) to get multiple reversed samples, which are then passed through the classifier, followed by majority voting of inferred labels to make the final prediction. This design of using multiple runs of denoising is informed by our theoretical analysis of the conditional distribution of the reversed sample. Specifically, when the data density of a clean sample is high, its conditional density under the reverse process in a diffusion model is also high; thus sampling from the latter conditional distribution can purify the adversarial example and return the corresponding clean sample with a high probability. By using the highest density point in the conditional distribution as the reversed sample, we identify the robust region of a given instance under the diffusion model's reverse process. We show that this robust region is a union of multiple convex sets, and is potentially much larger than the robust regions identified in previous works. In practice, DensePure can approximate the label of the high density region in the conditional distribution so that it can enhance certified robustness. We conduct extensive experiments to demonstrate the effectiveness of DensePure by evaluating its certified robustness given a standard model via randomized smoothing. We show that DensePure is consistently better than existing methods on ImageNet, with 7% improvement on average.

****

Towards Understanding How Machines Can Learn Causal Overhypotheses

Eliza Kosoy,David Chan,Adrian Liu,Jasmine Collins,Bryanna Kaufmann,Sandy Huang,Jessica B Hamrick,John Canny,Nan Rosemary Ke,Alison Gopnik

Recent work in machine learning and cognitive science has suggested that understanding causal information is essential to the development of intelligence. One of the key challenges for current machine learning algorithms is modeling and understanding causal overhypotheses: transferable abstract hypotheses about sets of causal relationships. In contrast, even young children spontaneously learn causal overhypotheses, and use these to guide their exploration or to generalize to new situations. This has been demonstrated in a variety of cognitive science experiments using the “blicket detector” environment. We present a causal learning benchmark adapting the “blicket" environment for machine learning agents and evaluate a range of state-of-the-art methods in this environment. We find that although most agents have no problem learning causal structures seen during training, they are unable to learn causal overhypotheses from these experiences, and thus cannot generalize to new settings.

****

Grounding Graph Network Simulators using Physical Sensor Observations

Jonas Linkerhägner,Niklas Freymuth,Paul Maria Scheikl,Franziska Mathis-Ullrich,Gerhard Neumann

Physical simulations that accurately model reality are crucial for many engineering disciplines such as mechanical engineering and robotic motion planning. In recent years, learned Graph Network Simulators produced accurate mesh-based simulations while requiring only a fraction of the computational cost of traditional simulators. Yet, the resulting predictors are confined to learning from data gen erated by existing mesh-based simulators and thus cannot include real world sensory information such as point cloud data. As these predictors have to simulate complex physical systems from only an initial state, they exhibit a high error accumulation for long-term predictions. In this work, we integrate sensory information to ground Graph Network Simulators on real world observations. In particular, we predict the mesh state of deformable objects by utilizing point cloud data. The resulting model allows for accurate predictions over longer time horizons, even under uncertainties in the simulation, such as unknown material properties. Since point clouds are usually not available for every time step, especially in online settings, we employ an imputation-based model. The model can make use of such additional information only when provided, and resorts to a standard Graph Network Simulator, otherwise. We experimentally validate our approach on a suite of prediction tasks for mesh-based interactions between soft and rigid bodies. Our method results in utilization of additional point cloud information to accurately predict stable simulations where existing Graph Network Simulators fail.

****

Skill Decision Transformer

Shyam Sudhakaran,Sebastian Risi

Recent work has shown that Large Language Models (LLMs) can be incredibly effective for offline reinforcement learning (RL) by representing the traditional RL problem as a sequence modelling problem. However many of these methods only optimize for high returns, and may not extract much information from a diverse dataset of trajectories. Generalized Decision Transformers (GDTs) have shown that by utilizing future trajectory information, in the form of information statistics, can help extract more information from offline trajectory data. Building upon this, we propose Skill Decision Transformer (Skill DT). Skill DT draws inspiration from hindsight relabelling and skill discovery methods to discover a diverse set of \emph{primitive behaviors}, or skills. We show that Skill DT can not only perform offline state-marginal matching (SMM), but can discovery descriptive behaviors that can be easily sampled. Furthermore, we show that through purely reward-free optimization, Skill DT is still competitive with supervised offline RL approaches on the D4RL benchmark.

****

In-distribution and Out-of-distribution Generalization for Graph Neural Networks

Emmanuel Sales,Renjie Liao,Nick Harvey

Graph neural networks (GNNs) are models that allow learning with structured data of varying size. Despite their popularity, theoretical understanding of the generalization of GNNs is an under-explored topic. In this work, we expand the theoretical understanding of both in-distribution and out-of-distribution generalization of GNNs. Firstly, we improve upon the state-of-the-art PAC-Bayes (in-distribution) generalization bound primarily by reducing an exponential dependency on the node degree to a linear dependency. Secondly, utilizing tools from spectral graph theory, we prove some rigorous guarantees about the out-of-distribution (OOD) size generalization of GNNs, where graphs in the training set have different numbers of nodes and edges from those in the test set. To empirically verify our theoretical findings, we conduct experiments on both synthetic and real-world graph datasets. Our computed generalization gaps for the in-distribution case significantly improve the state-of-the-art PAC-Bayes results. For the OOD case, experiments on community classification tasks in large social networks show that GNNs achieve strong size generalization performance in cases guaranteed by our theory.

****

Where to Diffuse, How to Diffuse, and How to Get Back: Automated Learning for Multivariate Diffusions

Raghav Singhal,Mark Goldstein,Rajesh Ranganath

Diffusion-based generative models (DBGMs) perturb data to a target noise distribution and reverse this process to generate samples. The choice of noising process, or inference diffusion process, affects both likelihoods and sample quality. For example, extending the inference process with auxiliary variables leads to improved sample quality. While there are many such multivariate diffusions to ex plore, each new one requires significant model-specific analysis, hindering rapid prototyping and evaluation. In this work, we study Multivariate Diffusion Models (MDMs). For any number of auxiliary variables, we provide a recipe for maximizing a lower-bound on the MDMs likelihood without requiring any model-specific analysis. We then demonstrate how to parameterize the diffusion for a specified target noise distribution; these two points together enable optimizing the inference diffusion process. Optimizing the diffusion expands easy experimentation from just a few well-known processes to an automatic search over all linear diffusions. To demonstrate these ideas, we introduce two new specific diffusions as well as learn a diffusion process on the MNIST, CIFAR10, and ImageNet32 datasets. We show learned MDMs match or surpass bits-per-dims (BPDs) relative to fixed choices of diffusions for a given dataset and model architecture.

****

Contrastive Corpus Attribution for Explaining Representations

Chris Lin,Hugh Chen,Chanwoo Kim,Su-In Lee

Despite the widespread use of unsupervised models, very few methods are designed to explain them. Most explanation methods explain a scalar model output. However, unsupervised models output representation vectors, the elements of which are not good candidates to explain because they lack semantic meaning. To bridge this gap, recent works defined a scalar explanation output: a dot product-based similarity in the representation space to the sample being explained (i.e., an explicand). Although this enabled explanations of unsupervised models, the interpretation of this approach can still be opaque because similarity to the explicand's representation may not be meaningful to humans. To address this, we propose contrastive corpus similarity, a novel and semantically meaningful scalar explanation output based on a reference corpus and a contrasting foil set of samples. We demonstrate that contrastive corpus similarity is compatible with many post-hoc feature attribution methods to generate COntrastive COrpus Attributions (COCOA) and quantitatively verify that features important to the corpus are identified. We showcase the utility of COCOA in two ways: (i) we draw insights by explaining augmentations of the same image in a contrastive learning setting (SimCLR); and (ii) we perform zero-shot object localization by explaining the similarity of image representations to jointly learned text representations (CLIP).

****

The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices

Will Hawkins,Brent Mittelstadt

The technical progression of artificial intelligence (AI) research has been built on breakthroughs in fields such as computer science, statistics, and mathematics. However, in the past decade AI researchers have increasingly looked to the social sciences, turning to human interactions to solve the challenges of model development. Paying crowdsourcing workers to generate or curate data, or ‘data enrichment’, has become indispensable for many areas of AI research, from natural language processing to inverse reinforcement learning. Other fields that routinely interact with crowdsourcing workers, such as Psychology, have developed common governance requirements and norms to ensure research is undertaken ethically. This study explores how, and to what extent, comparable research ethics requirements and norms have developed for AI research and data enrichment. We focus on the approach taken by two leading AI conferences: ICLR and NeurIPS. In a longitudinal study of accepted papers, and a comparison with Springer journal articles and Psychology papers, this work finds that ICLR and NeurIPS have established protocols for human data collection which are inconsistently followed by authors. Whilst Psychology papers engaging with crowdsourcing workers frequently disclose ethics reviews, payment data, demographic data and other information, such disclosures are far less common in leading AI conferences despite similar guidance. The work concludes with hypotheses to explain these gaps in research ethics practices and considerations for its implications.

****

Spatio-temporal point processes with deep non-stationary kernels

Zheng Dong,Xiuyuan Cheng,Yao Xie

Point process data are becoming ubiquitous in modern applications, such as social networks, health care, and finance. Despite the powerful expressiveness of the popular recurrent neural network (RNN) models for point process data, they may not successfully capture sophisticated non-stationary dependencies in the data due to their recurrent structures. Another popular type of deep model for point process data is based on representing the influence kernel (rather than the intensity function) by neural networks. We take the latter approach and develop a new deep non-stationary influence kernel that can model non-stationary spatio-temporal point processes. The main idea is to approximate the influence kernel with a novel and general low-rank decomposition, enabling efficient representation through deep neural networks and computational efficiency and better performance. We also take a new approach to maintain the non-negativity constraint of the conditional intensity by introducing a log-barrier penalty. We demonstrate our proposed method's good performance and computational efficiency compared with the state-of-the-art on simulated and real data.

****

Federated Learning from Small Datasets

Michael Kamp,Jonas Fischer,Jilles Vreeken

Federated learning allows multiple parties to collaboratively train a joint model without having to share any local data. It enables applications of machine learning in settings where data is inherently distributed and undisclosable, such as in the medical domain. Joint training is usually achieved by aggregating local models. When local datasets are small, locally trained models can vary greatly from a globally good model. Bad local models can arbitrarily deteriorate the aggregate model quality, causing federating learning to fail in these settings. We propose a novel approach that avoids this problem by interleaving model aggregation and permutation steps. During a permutation step we redistribute local models across clients through the server, while preserving data privacy, to allow each local model to train on a daisy chain of local datasets. This enables successful training in data-sparse domains. Combined with model aggregation, this approach enables effective learning even if the local datasets are extremely small, while retaining the privacy benefits of federated learning.

****

Explainable Machine Learning Predictions for the Long-term Performance of Brain-Computer Interfaces

Morgan E Urdaneta,Nicole C Veit,Renae G Burke,Ian G Malone,Kaleb E Smith,Kevin J Otto

Brain computer interfaces (BCIs) can decode neural signals to control assistive technologies such as robotic limbs for people with paralysis. Neural recordings from intracortical microelectrodes offer the spatiotemporal resolution (e.g., sortable units) necessary for complex tasks, such as controlling a robotic arm with multiple degrees of freedom. However, the quality of these signals decays over time despite many attempts to prolong their longevity. This decrease in long-term performance limits the implementation of this potentially beneficial technology. Predicting whether a channel will have sortable units across time would mitigate this issue and increase the utility of these devices by reducing uncertainty, yet to-date, no such methods exist. Similarly, it would be useful to understand how variables like time post-implantation, electrochemical characteristics, and electrode design impact the long-term quality of these signals. Here, we obtained longitudinal neural recordings and electrochemical data from freely behaving rats implanted with a custom designed microelectrode array with varying site areas, shank positions, and site depths. This dataset was used to develop an explainable artificial intelligence pipeline that predicts with high accuracy the presence of sortable units on a given channel and elucidates the most important factors leading to these predictions. Our pipeline was able to predict whether a channel will be active with an AUC of 0.79 (95% C.I. 0.73–0.86) on unseen data. The most important features of the model were experimental subject, time post-implantation, and a channel’s previous spike metrics. Electrode site depth was the most important electrode design variable. Our results demonstrate the feasibility of implementing explainable artificial intelligence pipelines for longitudinal

BCI studies and support previous reports on how factors like time, inter-animal variability, and cortical depth impact long-term performance of BCIs. These results are an important step forward in improving efficient decoding performance and guiding device development, which stand to advance the field and benefit the lives of human BCI patients.

****

Effectively using public data in privacy preserving Machine learning

Milad Nasr,Saeed Mahloujifar,Xinyu Tang,Prateek Mittal,Amir Houmansadr

A key challenge towards differentially private machine learning is balancing the trade-off between privacy and utility.

A recent line of work has demonstrated that leveraging \emph{public data samples} can enhance the utility of DP-trained models (for the same privacy guarantees).

In this work, we show that public data can be used to improve utility in DP models significantly more than shown in recent works.

Towards this end, we introduce a modified DP-SGD algorithm that leverages public data during its training process.

Our technique uses public data in two complementary ways: (1) it uses generative models trained on public data to produce synthetic data that is effectively embedded in multiple steps of the training pipeline; (2) it uses a new gradient clipping mechanism (required for achieving differential privacy) which changes the \emph{origin} of gradient vectors using information inferred from available public and synthesized data.

Our experimental results demonstrate the effectiveness of our approach in improving the state-of-the-art in differentially private machine learning across multiple datasets, network architectures, and application domains.

Notably, we achieve a 75% accuracy on CIFAR10 when using only $2,000$ public images; this is \emph{significantly higher} than the state-of-the-art which is 68% for DP-SGD with the privacy budget of $\varepsilon=2,\delta=10^{-5}$ (given the same number of public data points).

****

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation

Lorenz Kuhn,Yarin Gal,Sebastian Farquhar

We introduce a method to measure uncertainty in large language models. For tasks like question answering, it is essential to know when we can trust the natural language outputs of foundation models. We show that measuring uncertainty in natural language is challenging because of "semantic equivalence"—different sentences can mean the same thing. To overcome these challenges we introduce semantic entropy—an entropy which incorporates linguistic invariances created by shared meanings. Our method is unsupervised, uses only a single model, and requires no modifications to off-the-shelf language models. In comprehensive ablation studies we show that the semantic entropy is more predictive of model accuracy on question answering data sets than comparable baselines.

****

Illusory Adversarial Attacks on Sequential Decision-Makers and Countermeasures

Tim Franzmeyer,Stephen Marcus McAleer,Joao F. Henriques,Philip Torr,Jakob Nicolaus Foerster,Adel Bibi,Christian Schroeder de Witt

Autonomous decision-making agents deployed in the real world need to be robust against possible adversarial attacks on sensory inputs. Existing work on adversarial attacks focuses on the notion of perceptual invariance popular in computer vision. We observe that such attacks can often be detected by victim agents, since they result in action-observation sequences that are not consistent with the dynamics of the environment. Furthermore, real-world agents, such as physical robots, commonly operate under human supervisors who are not susceptible to such attacks. We propose to instead focus on attacks that are statistically undetectable. Specifically, we propose illusory attacks, a novel class of adversarial attack that is consistent with the environment dynamics. We introduce a novel algorithm that can learn illusory attacks end-to-end. We empirically verify that our algorithm generates attacks that, in contrast to current methods, are undetectable

to both AI agents with an environment dynamics model, as well as to humans. Furthermore, we show that existing robustification approaches are relatively ineffective against illusory attacks. Our findings highlight the need to ensure that real-world AI, and human-AI, systems are designed to make it difficult to corrupt sensory observations in ways that are consistent with the environment dynamics.
****

Prompt Tuning with Prompt-aligned Gradient for Vision-Language Models
Beier Zhu,Yulei Niu,Yucheng Han,Yue Wu,Hanwang Zhang

Thanks to the large pre-trained vision-language models (VLMs) like CLIP, we can craft a zero-shot classifier by ``prompt'', e.g., using the model provided similarity measure between an image and the prompt sentence ``$\texttt{a photo of a [CLASS]}$'', as the confidence score of predicting the image is ``$\texttt{[CLASS]}$''. Therefore, prompt shows a great potential for fast adapting the VLMs to downstream tasks if we fine-tune the prompt-based similarity measure. However, we find a common failure that improper fine-tuning may not only undermine the prompt's inherent prediction for the task-related classes, but also for other classes in the VLM vocabulary. Existing methods still address this problem by using traditional anti-overfitting techniques such as early stopping and data augmentation, which lack a principled solution specific to prompt. We present Prompt-aligned Gradient, dubbed $\texttt{ProGrad}$, to prevent prompt tuning from forgetting the the general knowledge learned from VLMs. In particular, $\texttt{ProGrad}$ only updates the prompt whose gradient is aligned (or non-conflicting) to the ``general direction'', which is represented as the gradient of the KL loss of the pre-defined prompt prediction. Extensive experiments demonstrate the stronger few-shot generalization ability of $\texttt{ProGrad}$ over state-of-the-art prompt tuning methods. Codes are in Appendix.
****

Relative Behavioral Attributes: Filling the Gap between Symbolic Goal Specification and Reward Learning from Human Preferences
Lin Guan,Karthik Valmeekam,Subbarao Kambhampati

Generating complex behaviors that satisfy the preferences of non-expert users is a crucial requirement for AI agents. Interactive reward learning from trajectory comparisons (a.k.a. RLHF) is one way to allow non-expert users to convey complex objectives by expressing preferences over short clips of agent behaviors. Even though this parametric method can encode complex tacit knowledge present in the underlying tasks, it implicitly assumes that the human is unable to provide richer feedback than binary preference labels, leading to intolerably high feedback complexity and poor user experience. While providing a detailed symbolic closed-form specification of the objectives might be tempting, it is not always feasible even for an expert user. However, in most cases, humans are aware of how the agent should change its behavior along meaningful axes to fulfill their underlying purpose, even if they are not able to fully specify task objectives symbolically. Using this as motivation, we introduce the notion of Relative Behavioral Attributes, which allows the users to tweak the agent behavior through symbolic concepts (e.g., increasing the softness or speed of agents' movement). We propose two practical methods that can learn to model any kind of behavioral attributes from ordered behavior clips. We demonstrate the effectiveness of our methods on four tasks with nine different behavioral attributes, showing that once the attributes are learned, end users can produce desirable agent behaviors relatively effortlessly, by providing feedback just around ten times. This is over an order of magnitude less than that required by the popular learning-from-human-preferences baselines. The supplementary video and source code are available at: https://guansuns.github.io/pages/rba.
****

DINO as a von Mises-Fisher mixture model
Hariprasath Govindarajan,Per Sidén,Jacob Roll,Fredrik Lindsten

Self-distillation methods using Siamese networks are popular for self-supervised pre-training. DINO is one such method based on a cross-entropy loss between K-dimensional probability vectors, obtained by applying a softmax function to the dot product between representations and learnt prototypes. Given the fact that the learned representations are L^2-normalized, we show that DINO and its derivatives, such as iBOT, can be interpreted as a mixture model of von Mises-Fisher components. With this interpretation, DINO assumes equal precision for all components when the prototypes are also L^2-normalized. Using this insight we propose DINO-vMF, that adds appropriate normalization constants when computing the cluster assignment probabilities. Unlike DINO, DINO-vMF is stable also for the larger ViT-Base model with unnormalized prototypes. We show that the added flexibility of the mixture model is beneficial in terms of better image representations. The DINO-vMF pre-trained model consistently performs better than DINO on a range of downstream tasks. We obtain similar improvements for iBOT-vMF vs iBOT and thereby show the relevance of our proposed modification also for other methods derived from DINO.

****

Continuous Depth Recurrent Neural Differential Equations

Srinivas Anumasa,Geetakrishnasai Gunapati,P. K. Srijith

Recurrent neural networks (RNNs) have brought a lot of advancements in sequence labeling tasks and sequence data. However, their effectiveness is limited when the observations in the sequence are irregularly sampled, where the observations arrive at irregular time intervals. To address this, continuous time variants of the RNNs were introduced based on neural ordinary differential equations (NODE). They learn a better representation of the data using the continuous transformation of hidden states over time, taking into account the time interval between the observations. However, they are still limited in their capability as they use the discrete transformations and discrete number of layers (depth) over an input in the sequence to produce the output observation. We intend to address this limitation by proposing RNNs based on differential equations which model continuous transformations over depth and time to predict an output for a given input in the sequence. Specifically, we propose continuous depth recurrent neural differential equations (CDR-NDE) which generalizes RNN models by continuously evolving the hidden states in both the temporal and depth dimensions. CDR-NDE considers two separate differential equations over each of these dimensions and models the evolution in the temporal and depth directions alternatively. We also propose the CDR-NDE-heat model based on partial differential equations which treats the computation of hidden states as solving a heat equation over time. We demonstrate the effectiveness of the proposed models by comparing against the state-of-the-art RNN models on real world sequence modeling problems and datasets.

****

Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms

Saeed Mahloujifar,Alexandre Sablayrolles,Graham Cormode,Somesh Jha

Given a trained model and a data sample, membership-inference (MI) attacks predict whether the sample was in the model’s training set. A common counter- measure against MI attacks is to utilize differential privacy (DP) during model training to mask the presence of individual examples. While this use of DP is a principled approach to limit the efficacy of MI attacks, there is a gap between the bounds provided by DP and the empirical performance of MI attacks. In this paper, we derive bounds for the advantage of an adversary mounting a MI attack, and demonstrate tightness for the widely-used Gaussian mechanism. Our analysis answers an open problem in the field of differential privacy, namely the fact that membership inference is not 100% successful even for relatively high budgets ($\epsilon> 10$). Finally, using our analysis, we provide MI metrics for models trained on CIFAR10 dataset. To the best of our knowledge, our analysis provides the state-of-the-art membership inference bounds.

****

Advantage Constrained Proximal Policy Optimization in Multi-Agent Reinforcement Learning

Weifan Li

We explore the value-based method and policy gradient combination in multi-agent reinforcement learning (MARL). In value-based MARL, {\itshape{Individual-Global-Max}} (IGM) principle plays an important role, which maintains the consistency between joint and local action values. At the same time, IGM is difficult to guarantee in multi-agent policy gradient methods due to stochastic exploration and conflicting gradient directions. In this paper, we propose a novel multi-agent policy gradient algorithm called {\itshape{Advantage Constrained Proximal Policy Optimization}} (ACPPO). Based on {\itshape{multi-agent advantage decomposition lemma}}, ACPPO introduces an advantage network for each agent to estimate current local state-action advantage. The coefficient of each agent constrains the joint-action advantage according to the consistency of the estimated joint-action advantage and local advantage. Unlike previous policy gradient-based MARL algorithms, ACPPO does not need an extra sampled baseline to reduce variance. We evaluate the proposed methods for continuous matrix game and Multi-Agent MuJoCo tasks. Results show that ACPPO outperforms the baselines such as MAPPO, MADDPG, and HAPPO.

****

Scalable Batch-Mode Deep Bayesian Active Learning via Equivalence Class Annealing

Renyu Zhang,Aly A Khan,Robert L. Grossman,Yuxin Chen

Active learning has demonstrated data efficiency in many fields. Existing active learning algorithms, especially in the context of batch-mode deep Bayesian active models, rely heavily on the quality of uncertainty estimations of the model, and are often challenging to scale to large batches. In this paper, we propose Batch-BALanCe, a scalable batch-mode active learning algorithm, which combines insights from decision-theoretic active learning, combinatorial information measure, and diversity sampling. At its core, Batch-BALanCe relies on a novel decision-theoretic acquisition function that facilitates differentiation among different equivalence classes. Intuitively, each equivalence class consists of hypotheses (e.g., posterior samples of deep neural networks) with similar predictions, and Batch-BALanCe adaptively adjusts the size of the equivalence classes as learning progresses. To scale up the computation of queries to large batches, we further propose an efficient batch-mode acquisition procedure, which aims to maximize a novel combinatorial information measure defined through the acquisition function. We show that our algorithm can effectively handle realistic multi-class classification tasks, and achieves compelling performance on several benchmark datasets for active learning under both low- and large-batch regimes.

****

Neural multi-event forecasting on spatio-temporal point processes using probabilistically enriched transformers

Negar Erfanian,Santiago Segarra,Maarten V. de Hoop

Predicting discrete events in time and space has many scientific applications, such as predicting hazardous earthquakes and outbreaks of infectious diseases. History-dependent spatio-temporal Hawkes processes are often used to mathematically model these point events. However, previous approaches have faced numerous challenges, particularly when attempting to forecast multiple future events. In this work, we propose a new neural architecture for multi-event forecasting of spatio-temporal point processes, utilizing transformers, augmented with normalizing flows and probabilistic layers. Our network makes batched predictions of complex history-dependent spatio-temporal distributions of future discrete events, achieving state-of-the-art performance on a variety of benchmark datasets including the South California Earthquakes, Citibike, Covid19, and Hawkes synthetic Pinwheel datasets. More generally, we illustrate how our network can be applied to any dataset of discrete events with associated markers, even when no underlying physics is known.

****

Associative Memory Augmented Asynchronous Spatiotemporal Representation Learning for Event-based Perception

Uday Kamal,Saurabh Dash,Saibal Mukhopadhyay

We propose $\textit{EventFormer}$, a computationally efficient event-based representation learning framework for asynchronously processing event camera data. Ev entFormer treats sparse input events as a spatially unordered set and models their spatial interactions using self-attention mechanism. An associative memory-augmented recurrent module is used to correlate with the stored representation computed from past events. A memory addressing mechanism is proposed to store and retrieve the latent states only $\textit{where}$ these events occur and update them only $\textit{when}$ they occur. The representation learning shift from input space to the latent memory space resulting in reduced computation cost for processing each event. We show that EventFormer achieves 0.5$\%$ and 9$\%$ better accuracy with 30000$\times$ and 200$\times$ less computation compared to the state-of-the-art dense and event-based method, respectively, on event-based object recognition datasets.

****

Detecting Small Query Graphs in A Large Graph via Neural Subgraph Search
Yunsheng Bai,Derek Qiang Xu,Yizhou Sun,Wei Wang

Recent advances have shown the success of using reinforcement learning and search to solve NP-hard graph-related tasks, such as Traveling Salesman Optimization, Graph Edit Distance computation, etc. However, it remains unclear how one can efficiently and accurately detect the occurrences of a small query graph in a large target graph, which is a core operation in graph database search, biomedical analysis, social group finding, etc. This task is called Subgraph Matching which essentially performs subgraph isomorphism check between a query graph and a large target graph. One promising approach to this classical problem is the “learning-to-search” paradigm, where a reinforcement learning (RL) agent is designed with a learned policy to guide a search algorithm to quickly find the solution without any solved instances for supervision. However, for the specific task of Subgraph Matching, though the query graph is usually small given by the user as input, the target graph is often orders-of-magnitude larger. It poses challenges to the neural network design and can lead to solution and reward sparsity. In this paper, we propose NSUBS with two innovations to tackle the challenges: (1) A novel encoder-decoder neural network architecture to dynamically compute the matching information between the query and the target graphs at each search state; (2) A novel look-ahead loss function for training the policy network. Experiments on six large real-world target graphs show that NSUBS can significantly improve the subgraph matching performance.

****

Catastrophic overfitting is a bug but it is caused by features
Guillermo Ortiz-Jimenez,Pau de Jorge,Amartya Sanyal,Adel Bibi,Puneet K. Dokania,Pascal Frossard,Grégory Rogez,Philip Torr

Adversarial training (AT) is the de facto method to build robust neural networks, but it is computationally expensive. To overcome this, fast single-step attacks can be used, but doing so is prone to catastrophic overfitting (CO). This is when networks gain non-trivial robustness during the first stages of AT, but then reach a breaking point where they become vulnerable in just a few iterations. Although some works have succeeded at preventing CO, the different mechanisms that lead to this failure mode are still poorly understood. In this work, we study the onset of CO in single-step AT methods through controlled modifications of typical datasets of natural images. In particular, we show that CO can be induced when injecting the images with seemingly innocuous features that are very useful for non-robust classification but need to be combined with other features to obtain a robust classifier. This new perspective provides important insights into the mechanisms that lead to CO and improves our understanding of the general dynamics of adversarial training.

****

Semi-Parametric Inducing Point Networks and Neural Processes
Richa Rastogi,Yair Schiff,Alon Hacohen,Zhaozhi Li,Ian Lee,Yuntian Deng,Mert R. Sabuncu,Volodymyr Kuleshov

We introduce semi-parametric inducing point networks (SPIN), a general-purpose architecture that can query the training set at inference time in a compute-efficient manner. Semi-parametric architectures are typically more compact than parametric models, but their computational complexity is often quadratic. In contrast

, SPIN attains linear complexity via a cross-attention mechanism between datapoints inspired by inducing point methods. Querying large training sets can be particularly useful in meta-learning, as it unlocks additional training signal, but often exceeds the scaling limits of existing models. We use SPIN as the basis of the Inducing Point Neural Process, a probabilistic model which supports large contexts in meta-learning and achieves high accuracy where existing models fail. In our experiments, SPIN reduces memory requirements, improves accuracy across a range of meta-learning tasks, and improves state-of-the-art performance on an important practical problem, genotype imputation.

****

CADet: Fully Self-Supervised Anomaly Detection With Contrastive Learning

Charles Guille-Escuret,Pau Rodriguez,David Vazquez,Ioannis Mitliagkas,Joao Monteiro

Handling out-of-distribution (OOD) samples has become a major stake in the real-world deployment of machine learning systems. This work explores the application of self-supervised contrastive learning to the simultaneous detection of two types of OOD samples: unseen classes and adversarial perturbations. Since in practice the distribution of such samples is not known in advance, we do not assume access to OOD examples. We show that similarity functions trained with contrastive learning can be leveraged with the maximum mean discrepancy (MMD) two-sample test to verify whether two independent sets of samples are drawn from the same distribution. Inspired by this approach, we introduce CADet (Contrastive Anomaly Detection), a method based on image augmentations to perform anomaly detection on single samples. CADet compares favorably to adversarial detection methods to detect adversarially perturbed samples on ImageNet. Simultaneously, it achieves comparable performance to unseen label detection methods on two challenging benchmarks: ImageNet-O and iNaturalist. Additionally, CADet is fully self-supervised and requires neither labels for in-distribution samples nor access to out-of-distribution examples.

****

SPRINT: Scalable Semantic Policy Pre-training via Language Instruction Relabeling

Jesse Zhang,Karl Pertsch,Jiahui Zhang,Taewook Nam,Sung Ju Hwang,Xiang Ren,Joseph J Lim

We propose SPRINT, an approach for scalable offline policy pre-training based on natural language instructions. SPRINT pre-trains an agent’s policy to execute a diverse set of semantically meaningful skills that it can leverage to learn new tasks faster. Prior work on offline pre-training required tedious manual definition of pre-training tasks or learned semantically meaningless skills via random goal-reaching. Instead, our approach SPRINT (Scalable Pre-training via Relabeling Language INsTructions) leverages natural language instruction labels on offline agent experience, collected at scale (e.g., via crowd-sourcing), to define a rich set of tasks with minimal human effort. Furthermore, by using natural language to define tasks, SPRINT can use pre-trained large language models to automatically expand the initial task set. By relabeling and aggregating task instructions, even across multiple training trajectories, we can learn a large set of new skills during pre-training. In experiments using a realistic household simulator, we show that agents pre-trained with SPRINT learn new long-horizon household tasks substantially faster than with previous pre-training approaches.

****

SMART: Self-supervised Multi-task pretrAining with contRol Transformers

Yanchao Sun,Shuang Ma,Ratnesh Madaan,Rogerio Bonatti,Furong Huang,Ashish Kapoor

Self-supervised pretraining has been extensively studied in language and vision domains, where a unified model can be easily adapted to various downstream tasks by pretraining representations without explicit labels. When it comes to sequential decision-making tasks, however, it is difficult to properly design such a pretraining approach that can cope with both high-dimensional perceptual information and the complexity of sequential control over long interaction horizons. The challenge becomes combinatorially more complex if we want to pretrain representations amenable to a large variety of tasks. To tackle this problem, in this wor k, we formulate a general pretraining-finetuning pipeline for sequential decision making, under which we propose a generic pretraining framework \textit{Self-supervised Multi-task pretrAining with contRol Transformer (SMART)}. By systematically investigating pretraining regimes, we carefully design a Control Transformer (CT) coupled with a novel control-centric pretraining objective in a self-supervised manner. SMART encourages the representation to capture the common essential information relevant to short-term control and long-term control, which is transferrable across tasks. We show by extensive experiments in DeepMind Control Suite that SMART significantly improves the learning efficiency among seen and unseen downstream tasks and domains under different learning scenarios including Imitation Learning (IL) and Reinforcement Learning (RL). Benefiting from the proposed control-centric objective, SMART is resilient to distribution shift between pretraining and finetuning, and even works well with low-quality pretraining datasets that are randomly collected. The codebase, pretrained models and datasets are provided at https://github.com/microsoft/smart.

****

Evaluation of Active Feature Acquisition Methods under Missing Data

Henrik von Kleist,Alireza Zamanian,Ilya Shpitser,Narges Ahmidi

Machine learning (ML) methods generally assume the full set of features are available at no cost. If the acquisition of a certain feature is costly at run-time, one might want to balance the acquisition cost and the predictive value of the feature for the ML task. The task of training an AI agent to decide which features are necessary to be acquired is called active feature acquisition (AFA). Current AFA methods, however, are challenged when the AFA agent has to be trained/tested with datasets that contain missing data. We formulate, for the first time, the problem of active feature acquisition performance evaluation (AFAPE) under missing data, i.e. the problem of adjusting for the inevitable missingness distribution shift between train/test time and run-time. We first propose a new causal graph, the AFA graph, that characterizes the AFAPE problem as an intervention on the reinforcement learning environment used to train AFA agents. Here, we discuss that for handling missing data in AFAPE, the conventional approaches (off-policy reinforcement learning, blocked feature acquisitions, imputation and inverse probability weighting (IPW)) often lead to biased results or are data inefficient. We then propose active feature acquisition importance sampling (AFAIS), a novel estimator that is more data efficient than IPW. We demonstrate the detrimental conclusions to which biased estimators can lead as well as the high data efficiency of AFAIS in multiple experiments using simulated and real-world data under induced MCAR, MAR and MNAR missingness.

****

DAG Learning on the Permutahedron

Valentina Zantedeschi,Luca Franceschi,Jean Kaddour,Matt Kusner,Vlad Niculae

We propose a continuous optimization framework for discovering a latent directed acyclic graph (DAG) from observational data. Our approach optimizes over the polytope of permutation vectors, the so-called Permutahedron, to learn a topological ordering. Edges can be optimized jointly, or learned conditional on the ordering via a non-differentiable subroutine. Compared to existing continuous optimization approaches our formulation has a number of advantages including: 1. validity: optimizes over exact DAGs as opposed to other relaxations optimizing approximate DAGs; 2. modularity: accommodates any edge-optimization procedure, edge structural parameterization, and optimization loss; 3. end-to-end: either alternately iterates between node-ordering and edge-optimization, or optimizes them jointly; We demonstrate, on real-world data problems in protein-signaling and transcriptional network discovery, that our approach lies on the Pareto frontier of two key metrics, the SID and SHD.

****

Explicitly Minimizing the Blur Error of Variational Autoencoders

Gustav Bredell,Kyriakos Flouris,Krishna Chaitanya,Ertunc Erdil,Ender Konukoglu

Variational autoencoders (VAEs) are powerful generative modelling methods, however they suffer from blurry generated samples and reconstructions compared to the images they have been trained on. Significant research effort has been spent to increase the generative capabilities by creating more flexible models but often flexibility comes at the cost of higher complexity and computational cost. Several works have focused on altering the reconstruction term of the evidence lower bound (ELBO), however, often at the expense of losing the mathematical link to maximizing the likelihood of the samples under the modeled distribution. Here we propose a new formulation of the reconstruction term for the VAE that specifically penalizes the generation of blurry images while at the same time still maximizing the ELBO under the modeled distribution.

We show the potential of the proposed loss on three different data sets, where it outperforms several recently proposed reconstruction losses for VAEs.

****

GraphEditor: An Efficient Graph Representation Learning and Unlearning Approach

Weilin Cong,Mehrdad Mahdavi

As graph representation learning has received much attention due to its widespread applications, removing the effect of a specific node from the pre-trained graph representation learning model due to privacy concerns has become equally important.

However, due to the dependency between nodes in the graph, graph representation unlearning is notoriously challenging and still remains less well explored. To fill in this gap, we propose \textsc{GraphEditor}, an efficient graph representation \textit{learning} and \textit{unlearning} approach that supports node/edge deletion, node/edge addition, and node feature update. Compared to existing unlearning approaches, \textsc{GraphEditor} requires neither retraining from scratch nor of all data presented during unlearning, which is beneficial for the settings that not all the training data are available to retrain. Besides, since \textsc{GraphEditor} is exact unlearning, the removal of all the information associated with the deleted nodes/edges can be guaranteed. Empirical results on real-world datasets illustrate the effectiveness of \textsc{GraphEditor} for both node and edge unlearning tasks.

****

3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction

Jiaqi Guan,Wesley Wei Qian,Xingang Peng,Yufeng Su,Jian Peng,Jianzhu Ma

Rich data and powerful machine learning models allow us to design drugs for a specific protein target <em>in silico</em>. Recently, the inclusion of 3D structures during targeted drug design shows superior performance to other target-free models as the atomic interaction in the 3D space is explicitly modeled. However, current 3D target-aware models either rely on the voxelized atom densities or the autoregressive sampling process, which are not equivariant to rotation or easily violate geometric constraints resulting in unrealistic structures. In this work, we develop a 3D equivariant diffusion model to solve the above challenges. To achieve target-aware molecule design, our method learns a joint generative process of both continuous atom coordinates and categorical atom types with a SE(3)-equivariant network. Moreover, we show that our model can serve as an unsupervised feature extractor to estimate the binding affinity under proper parameterization, which provides an effective way for drug screening. To evaluate our model, we propose a comprehensive framework to evaluate the quality of sampled molecules from different dimensions. Empirical studies show our model could generate molecules with more realistic 3D structures and better affinities towards the protein targets, and improve binding affinity ranking and prediction without retraining.

****

PGASL: Predictive and Generative Adversarial Semi-supervised Learning for imbalanced data

Ru Huang,Yunlong Wang,Ruoxin Li,Wei Huang,Emily Zhao,Yilian Yuan

Modern machine learning techniques often suffer from class imbalance where only a small amount of data is available for minority classes. Classifiers trained on an imbalanced dataset, although have high accuracy on majority classes, can perform poorly on minority classes. This is problematic when minority classes are also important. Generative Adversarial Networks (GANs) have been proposed for gen erating artificial minority examples to balance the training. We propose a class-imbalanced semi-supervised learning algorithm PGASL which can be efficiently trained on unlabeled and class-imbalanced data. In this work, we use a predictive network which is trained adversarially for the discriminator to correct predictions on the unlabeled dataset. Experiments on text datasets show that PGASL outperforms state-of-the-art class-imbalanced learning algorithms by including both predictive network and generator.

****

Towards a More Rigorous Science of Blindspot Discovery in Image Models
Gregory Plumb,Nari Johnson,Angel Cabrera,Ameet Talwalkar
A growing body of work studies Blindspot Discovery Methods (BDMs): methods for finding semantically meaningful subsets of the data where an image classifier performs significantly worse, without making strong assumptions. Motivated by observed gaps in prior work, we introduce a new framework for evaluating BDMs, SpotCheck, that uses synthetic image datasets to train models with known blindspots and a new BDM, PlaneSpot, that uses a 2D image representation. We use SpotCheck to run controlled experiments that identify factors that influence BDM performance (e.g., the number of blindspot in a model) and show that PlaneSpot outperforms existing BDMs. Importantly, we validate these findings using real data. Overall, we hope that the methodology and analyses presented in this work will serve as a guide for future work on blindspot discovery.

****

How gradient estimator variance and bias impact learning in neural networks
Arna Ghosh,Yuhan Helena Liu,Guillaume Lajoie,Konrad Kording,Blake Aaron Richards
There is growing interest in understanding how real brains may approximate gradients and how gradients can be used to train neuromorphic chips. However, neither real brains nor neuromorphic chips can perfectly follow the loss gradient, so parameter updates would necessarily use gradient estimators that have some variance and/or bias. Therefore, there is a need to understand better how variance and bias in gradient estimators impact learning dependent on network and task properties. Here, we show that variance and bias can impair learning on the training data, but some degree of variance and bias in a gradient estimator can be beneficial for generalization. We find that the ideal amount of variance and bias in a gradient estimator are dependent on several properties of the network and task: the size and activity sparsity of the network, the norm of the gradient, and the curvature of the loss landscape. As such, whether considering biologically-plausible learning algorithms or algorithms for training neuromorphic chips, researchers can analyze these properties to determine whether their approximation to gradient descent will be effective for learning given their network and task properties.

****

Automatically Auditing Large Language Models via Discrete Optimization
Erik Jones,Anca Dragan,Aditi Raghunathan,Jacob Steinhardt
Auditing large language models for unexpected behaviors is critical to preempt catastrophic deployments, yet remains challenging. In this work, we cast auditing as a discrete optimization problem, where we automatically search for input-output pairs that match a desired target behavior. For example, we might aim to find non-toxic input that starts with ``Barack Obama'' and maps to a toxic output. Our optimization problem is difficult to solve as the set of feasible points is sparse, the space is discrete, and the language models we audit are non-linear and high-dimensional. To combat these challenges, we introduce a discrete optimization algorithm, ARCA, that is tailored to autoregressive language models. We demonstrate how our approach can: uncover derogatory completions about celebrities (e.g. ``Barack Obama is a legalized unborn'' $\rightarrow$ ``child murderer'), produce French inputs that complete to English outputs, and find inputs that generate a specific name. Our work offers a promising new tool to uncover models' failure-modes before deployment. $\textbf{Trigger Warning: This paper contains model behavior that can be offensive in nature.}$

****

Do We Really Need Complicated Model Architectures For Temporal Networks?

Weilin Cong,Si Zhang,Jian Kang,Baichuan Yuan,Hao Wu,Xin Zhou,Hanghang Tong,Mehrdad Mahdavi

Recurrent neural network (RNN) and self-attention mechanism (SAM) are the de facto methods to extract spatial-temporal information for temporal graph learning. Interestingly, we found that although both RNN and SAM could lead to a good performance, in practice neither of them is always necessary. In this paper, we propose GraphMixer, a conceptually and technically simple architecture that consists of three components: (1) a link-encoder that is only based on multi-layer perceptrons (MLP) to summarize the information from temporal links, (2) a node-encoder that is only based on neighbor mean-pooling to summarize node information, and (3) an MLP-based link classifier that performs link prediction based on the outputs of the encoders. Despite its simplicity, GraphMixer attains an outstanding performance on temporal link prediction benchmarks with faster convergence and better generalization performance. These results motivate us to rethink the importance of simpler model architecture.

****

On the System-Level Effectiveness of Physical Object-Hiding Adversarial Attack in Autonomous Driving

Ningfei Wang,Yunpeng Luo,TAKAMI SATO,Kaidi Xu,Qi Alfred Chen

In Autonomous Driving (AD) systems, perception is crucial for both security and safety. Among the different attacks on AD perception, the physical object-hiding adversarial attacks are especially severe due to their direct impact on road safety.

However, we find that all existing works so far only evaluate their attack effect at the targeted AI component level, without any evaluation \textit{at the system level}, i.e., with the entire system semantics and context such as the full AD system pipeline and closed-loop control. This thus inevitably raise a critical research question: can these existing research efforts actually effectively achieve the desired system-level attack effects (e.g., causing vehicle collisions, traffic rule violations, etc.) in the real-world AD system context?

In the paper, we thus perform the first measurement study on whether and how effective the existing designs can lead to system-level effects, where we take the STOP sign-hiding attack as our target. Our evaluation results show that all the representative prior works cannot achieve any system-level effect in a classical closed-loop AD setup in road speeds controlled by common STOP signs. With that, we then point out two limitation hypotheses that appear in all existing works: 1) the unpractical STOP sign size distribution in pixel sampling, and 2) missing particular consideration in system-critical attack range. Experimental results demonstrate that after overcoming these two limitations, the system-level effects can be further improved, i.e., the violation rate can increase around 70\%.

****

Data Feedback Loops: Model-driven Amplification of Dataset Biases

Rohan Taori,Tatsunori Hashimoto

Datasets scraped from the internet have been critical to large-scale machine learning. Yet, its success puts the utility of future internet-derived datasets at potential risk, as model outputs begin to replace human annotations as a source of supervision. In this work, we formalize a system where interactions with one model are recorded as history and scraped as training data in the future. We then analyze its stability over time by tracking changes to a test-time bias statistic (e.g. gender bias of model predictions). We find that the degree of bias amplification is closely linked to whether the model’s outputs behave like samples from the training distribution, a behavior which we characterize and define as consistent calibration. Experiments in three conditional prediction scenarios – image classification, visual role-labeling, and language generation – demonstrate that models that exhibit a sampling-like behavior are more calibrated and thus more stable. Based on this insight, we propose an intervention to help calibrate and stabilize unstable feedback systems.

****

A 2-parameter Persistence Layer for Learning

Cheng Xin,Soham Mukherjee,Shreyas N. Samaga,Tamal K. Dey

1-parameter persistent homology, a cornerstone in Topological Data Analysis (TDA), studies the evolution of topological features such as cycle basis hidden in data. It has found its application in strengthening the representation power of deep learning models like Graph Neural Networks (GNN). To enrich the representations of topological features, here we propose to study 2-parameter persistence modules induced by bi-filtration functions. In order to incorporate these representations into machine learning models, we introduce a novel vectorization on 2-parameter persistence modules called Generalized Rank Invariant Landscape {\textsc{Gril}}. We show that this vector representation is stable and differentiable with respect to underlying filtration functions and can be easily integrated into machine learning models to augment encoding topological features. We present an algorithm to compute the vectorization and its gradients. We also test our methods on synthetic graph datasets and compare the results with some popular graph neural networks.

****

Is Conditional Generative Modeling all you need for Decision Making?

Anurag Ajay,Yilun Du,Abhi Gupta,Joshua B. Tenenbaum,Tommi S. Jaakkola,Pulkit Agrawal

Recent improvements in conditional generative modeling have made it possible to generate high-quality images from language descriptions alone. We investigate whether these methods can directly address the problem of sequential decision-making. We view decision-making not through the lens of reinforcement learning (RL), but rather through conditional generative modeling. To our surprise, we find that our formulation leads to policies that can outperform existing offline RL approaches across standard benchmarks. By modeling a policy as a return-conditional generative model, we avoid the need for dynamic programming and subsequently eliminate many of the complexities that come with traditional offline RL. We further demonstrate the advantages of modeling policies as conditional generative models by considering two other conditioning variables: constraints and skills. Conditioning on a single constraint or skill during training leads to behaviors at test-time that can satisfy several constraints together or demonstrate a composition of skills. Our results illustrate that conditional generative modeling is a powerful tool for decision-making.

****

META-STORM: Generalized Fully-Adaptive Variance Reduced SGD for Unbounded Functions

Zijian Liu,Ta Duy Nguyen,Thien Hang Nguyen,Alina Ene,Huy Nguyen

We study the application of variance reduction (VR) techniques to general non-convex stochastic optimization problems. In this setting, the recent work STORM (Cutkosky & Orabona, 2019) overcomes the drawback of having to compute gradients of “mega-batches” that earlier VR methods rely on. There, STORM utilizes recursive momentum to achieve the VR effect and is then later made fully adaptive in STORM+ (Levy et al., 2021), where full-adaptivity removes the requirement for obtaining certain problem-specific parameters such as the smoothness of the objective and bounds on the variance and norm of the stochastic gradients in order to set the step size. However, STORM+ crucially relies on the assumption that the function values are bounded, excluding a large class of useful functions. In this work, we propose META-STORM, a generalized framework of STORM+ that removes this bounded function values assumption while still attaining the optimal convergence rate for non-convex optimization. META-STORM not only maintains full-adaptivity, removing the need to obtain problem specific parameters, but also improves the convergence rate’s dependency on the problem parameters. Furthermore, META-STORM can utilize a large range of parameter settings that subsumes previous methods allowing for more flexibility in a wider range of settings. Finally, we demonstrate the effectiveness of META-STORM through experiments across common deep learning tasks. Our algorithm improves upon the previous work STORM+ and is competitive with widely used algorithms after the addition of per-coordinate update and exponential moving average heuristics.

****

TEMPERA: Test-Time Prompt Editing via Reinforcement Learning

Tianjun Zhang,Xuezhi Wang,Denny Zhou,Dale Schuurmans,Joseph E. Gonzalez

Careful prompt design is critical to the use of large language models in zero-shot or few-shot learning. As a consequence, there is a growing interest in automated methods to design optimal prompts. In this work, we propose Test-time Prompt Editing using Reinforcement learning (TEMPERA). In contrast to prior prompt generation methods, TEMPERA can efficiently leverage prior knowledge, is adaptive to different queries and provides an interpretable prompt for every query. To achieve this, we design a novel action space that allows flexible editing of the initial prompts covering a wide set of commonly-used components like instructions, few-shot exemplars, and verbalizers. The proposed method achieves significant gains compared with recent SoTA approaches like prompt tuning, AutoPrompt, and RLPrompt, across a variety of tasks including sentiment analysis, topic classification, natural language inference, and reading comprehension. Our method achieves 5.33x on average improvement in sample efficiency when compared to the traditional fine-tuning methods.

****

Combining pretrained speech and text encoders for spoken language processing

Karan Singla,Daniel Pressel,Ryan Price,Mahnoosh Mehrabani,Yeon-Jun Kim,Srinivas Bangalore

Spoken Language Processing tasks that extract information from speech signal, have the advantage of using both speech and text modalities. In this paper, we propose to combine pretrained speech and text encoders via cross-attention, and we show the application of the proposed architecture in multiple spoken language processing systems. Our results indicate that it's more efficient to re-purpose previously trained independent modality encoders and learn only cross-attention from scratch. This resultant architecture captures both acoustic and lexical information, and performs text tagging while attending to speech encoder for improved results. We use compact pretrained speech and text encoder which are resource efficient and can be trained on a single consumer GPU card.

****

A Large Scale Sample Complexity Analysis of Neural Policies in the Low-Data Regime

Ezgi Korkmaz

The progress in reinforcement learning algorithm development is at one of its highest points starting from the initial study that enabled sequential decision making from high-dimensional observations. Currently, deep reinforcement learning research has had quite recent breakthroughs from learning without the presence of rewards to learning functioning policies without even knowing the rules of the game. In our paper we focus on the trends currently used in deep reinforcement learning algorithm development in the low-data regime. We theoretically show that the performance profiles of the algorithms developed for the high-data regime do not transfer to the low-data regime in the same order. We conduct extensive experiments in the Arcade Learning Environment and our results demonstrate that the baseline algorithms perform significantly better in the low-data regime compared to the set of algorithms that were initially designed and compared in the large-data region.

****

Evaluating Representations with Readout Model Switching

Yazhe Li,Jorg Bornschein,Marcus Hutter

Although much of the success of Deep Learning builds on learning good representations, a rigorous method to evaluate their quality is lacking. In this paper, we treat the evaluation of representations as a model selection problem and propose to use the Minimum Description Length (MDL) principle to devise an evaluation metric. Contrary to the established practice of limiting the capacity of the readout model, we design a hybrid discrete and continuous-valued model space for the readout models and employ a switching strategy to combine their predictions. The MDL score takes model complexity, as well as data efficiency into account. As a result, the most appropriate model for the specific task and representation w ill be chosen, making it a unified measure for comparison. The proposed metric can be efficiently computed with an online method and we present results for pre-trained vision encoders of various architectures (ResNet and ViT) and objective functions (supervised and self-supervised) on a range of downstream tasks. We compare our methods with accuracy-based approaches and show that the latter are inconsistent when multiple readout models are used. Finally, we discuss important properties revealed by our evaluations such as model scaling, preferred readout model, and data efficiency.

****

Provable Defense Against Geometric Transformations

Rem Yang,Jacob Laurel,Sasa Misailovic,Gagandeep Singh

Geometric image transformations that arise in the real world, such as scaling and rotation, have been shown to easily deceive deep neural networks (DNNs). Hence, training DNNs to be certifiably robust to these perturbations is critical. However, no prior work has been able to incorporate the objective of deterministic certified robustness against geometric transformations into the training procedure, as existing verifiers are exceedingly slow. To address these challenges, we propose the first provable defense for deterministic certified geometric robustness. Our framework leverages a novel GPU-optimized verifier that can certify images between 60$\times$ to 42,600$\times$ faster than existing geometric robustness verifiers, and thus unlike existing works, is fast enough for use in training. Across multiple datasets, our results show that networks trained via our framework consistently achieve state-of-the-art deterministic certified geometric robustness and clean accuracy. Furthermore, for the first time, we verify the geometric robustness of a neural network for the challenging, real-world setting of autonomous driving.

****

Augmentation with Projection: Towards an Effective and Efficient Data Augmentation Paradigm for Distillation

Ziqi Wang,Yuexin Wu,Frederick Liu,Daogao Liu,Le Hou,Hongkun Yu,Jing Li,Heng Ji

Knowledge distillation is one of the primary methods of transferring knowledge from large to small models. However, it requires massive task-specific data, which may not be plausible in many real-world applications. Data augmentation methods such as representation interpolation, token replacement, or augmentation with models are applied to tackle this problem. However, these data augmentation methods either potentially cause shifts in decision boundaries (representation interpolation), are not expressive enough (token replacement), or introduce too much computational overhead (augmentation with models). To this end, we propose AugPro (Augmentation with Projection), an effective and efficient data augmentation method for distillation. Our method builds on top of representation interpolation augmentation methods to maintain the diversity of expressions and converts the augmented data to tokens to avoid shifting decision boundaries. It uses simple operations that come with little computational overhead. The results on multiple GLUE tasks show that our methods can improve distillation performance by a large margin at a low time cost.

****

Pseudoinverse-Guided Diffusion Models for Inverse Problems

Jiaming Song,Arash Vahdat,Morteza Mardani,Jan Kautz

Diffusion models have become competitive candidates for solving various inverse problems. Models trained for specific inverse problems work well but are limited to their particular use cases, whereas methods that use problem-agnostic models are general but often perform worse empirically. To address this dilemma, we introduce Pseudoinverse-guided Diffusion Models (ΠGDM), an approach that uses problem-agnostic models to close the gap in performance. ΠGDM directly estimates conditional scores from the measurement model of the inverse problem without additional training. It can address inverse problems with noisy, non-linear, or even non-differentiable measurements, in contrast to many existing approaches that are limited to noiseless linear ones. We illustrate the empirical effectiveness of ΠGDM on several image restoration tasks, including super-resolution, inpainting and JPEG restoration. On ImageNet, ΠGDM is competitive with stat e-of-the-art diffusion models trained on specific tasks, and is the first to achieve this with problem-agnostic diffusion models. ΠGDM can also solve a wider set of inverse problems where the measurement processes are composed of several simpler ones.

****

Autoregressive Diffusion Model for Graph Generation

Lingkai Kong,Jiaming Cui,Haotian Sun,Yuchen Zhuang,B. Aditya Prakash,Chao Zhang

Diffusion-based graph generative models have recently obtained promising results for graph generation. However, existing diffusion-based graph generative models are all one-shot generative models that apply Gaussian diffusion in the dequantized adjacency matrix space. Such a strategy can suffer from difficulty in model training, slow sampling speed, and incapability of incorporating constraints. We propose an \emph{autoregressive diffusion} model for graph generation. Unlike existing methods, we define a node-absorbing diffusion process that operates directly in the discrete graph space. For forward diffusion, we design a \emph{diffusion ordering network}, which learns an optimal node absorbing ordering from graph topology. For reverse generation, we design a \emph{denoising network} that uses the reverse node ordering to efficiently reconstruct the graph by predicting one row of the adjacency matrix at a time. Based on permutation invariance of graph generation, we show that the two networks can be jointly trained by optimizing a simple lower bound of data likelihood. Our experiments on six diverse datasets show that our model achieves better or comparable generation performance with previous state-of-the-art, and meanwhile enjoys fast generation speed.

****

Self-supervised video pretraining yields strong image representations

Nikhil Parthasarathy,S. M. Ali Eslami,Joao Carreira,Olivier J Henaff

Videos contain far more information than still images, and hold the potential for learning rich representations of the visual world. Yet, pretraining on image datasets has remained the dominant paradigm for learning representations that capture spatial information and previous attempts at video pretraining have fallen short on image understanding tasks. In this work we revisit self-supervised learning of image representations from the dynamic evolution of video frames. To that end, we propose a dataset curation procedure that addresses the domain mismatch between video and image datasets, and develop a contrastive learning framework which handles the complex transformations present in natural videos. This simple paradigm for distilling knowledge from videos to image representations, called VITO, performs surprisingly well on a variety of image-based transfer learning tasks. For the first time, our video-pretrained model closes the gap with ImageNet pretraining on semantic segmentation on PASCAL and ADE20k and object detection on COCO and LVIS, raising the possibility of video-pretraining becoming the new default for learning image representations.

****

Planning with Sequence Models through Iterative Energy Minimization

Hongyi Chen,Yilun Du,Yiye Chen,Joshua B. Tenenbaum,Patricio A. Vela

Recent works have shown that language modeling can be effectively used to train reinforcement learning (RL) policies. However, the success of applying existing language models to planning, in which we wish to obtain a trajectory of actions to reach some goal, is less straightforward. The typical autoregressive generation procedures of language models preclude sequential refinement of earlier steps, which limits the effectiveness of a predicted plan. In this paper, we suggest an approach towards integrating planning with language models based on the idea of iterative energy minimization, and illustrate how such a procedure leads to improved RL performance across different tasks. We train a masked language model to capture an implicit energy function over trajectories of actions, and formulate planning as finding a trajectory of actions with minimum energy. We illustrate how this procedure enables improved performance over recent approaches across BabyAI and Atari environments. We further demonstrate unique benefits of our iterative optimization procedure, involving new task generalization, test-time constraints adaptation, and the ability to compose plans together. Project webpage: https://hychen-naza.github.io/projects/LEAP/index.html

****

Verifying the Union of Manifolds Hypothesis for Image Data

Bradley CA Brown,Anthony L. Caterini,Brendan Leigh Ross,Jesse C Cresswell,Gabriel Loaiza-Ganem

Deep learning has had tremendous success at learning low-dimensional representations of high-dimensional data. This success would be impossible if there was no hidden low-dimensional structure in data of interest; this existence is posited by the manifold hypothesis, which states that the data lies on an unknown manifold of low intrinsic dimension. In this paper, we argue that this hypothesis does not properly capture the low-dimensional structure typically present in image data. Assuming that data lies on a single manifold implies intrinsic dimension is identical across the entire data space, and does not allow for subregions of this space to have a different number of factors of variation. To address this deficiency, we consider the union of manifolds hypothesis, which states that data lies on a disjoint union of manifolds of varying intrinsic dimensions. We empirically verify this hypothesis on commonly-used image datasets, finding that indeed, observed data lies on a disconnected set and that intrinsic dimension is not constant. We also provide insights into the implications of the union of manifolds hypothesis in deep learning, both supervised and unsupervised, showing that designing models with an inductive bias for this structure improves performance across classification and generative modelling tasks. Our code is available at https://github.com/layer6ai-labs/UoMH.

****

Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations

Polina Kirichenko,Pavel Izmailov,Andrew Gordon Wilson

Neural network classifiers can largely rely on simple spurious features, such as image backgrounds, to make predictions. However, even in these cases, we show that they still often learn core features associated with the desired attributes of the data, contrary to recent findings. Inspired by this insight, we demonstrate that simple last layer retraining can match or outperform state-of-the-art approaches on spurious correlation benchmarks, but with profoundly lower complexity and computational expenses. Moreover, we show that last layer retraining on large ImageNet-trained models can also significantly reduce reliance on background and texture information, improving robustness to covariate shift, after only minutes of training on a single GPU.

****

Progressive Data Dropout: An Adaptive Training Strategy for Large-Scale Supervised Learning

David Lee Patrick,Zachary Seligman,Amanda S Fernandez

Common training strategies for deep neural networks are computationally expensive, continuing to redundantly train and evaluate on classes already well-understood by the model. A common strategy to diminish this cost is to reduce data used in training, however this often comes at the expense of the model's accuracy or an additional computational cost in training. We propose progressive data dropout (PDD), an adaptive training strategy which performs class-level data dropout from the training set as the network develops an understanding for each class. Our experiments on large-scale image classification demonstrate PDD reduces the total number of datapoints needed to train the network by a factor of 10, reducing the overall training time without significantly impacting accuracy or modifying the model architecture. We additionally demonstrate improvements via experiments and ablations on computer vision benchmarks, including MNIST, Fashion-MNIST, SVHN, CIFAR, and ImageNet datasets.

****

Error Sensitivity Modulation based Experience Replay: Mitigating Abrupt Representation Drift in Continual Learning

Fahad Sarfraz,Elahe Arani,Bahram Zonooz

Humans excel at lifelong learning, as the brain has evolved to be robust to distribution shifts and noise in our ever-changing environment. Deep neural networks (DNNs), however, exhibit catastrophic forgetting and the learned representations drift drastically as they encounter a new task. This alludes to a different er ror-based learning mechanism in the brain. Unlike DNNs, where learning scales linearly with the magnitude of the error, the sensitivity to errors in the brain decreases as a function of their magnitude. To this end, we propose "ESMER" which employs a principled mechanism to modulate error sensitivity in a dual-memory rehearsal-based system. Concretely, it maintains a memory of past errors and uses it to modify the learning dynamics so that the model learns more from small consistent errors compared to large sudden errors. We also propose "Error-Sensitive Reservoir Sampling" to maintain episodic memory, which leverages the error history to pre-select low-loss samples as candidates for the buffer, which are better suited for retaining information. Empirical results show that ESMER effectively reduces forgetting and abrupt drift in representations at the task boundary by gradually adapting to the new task while consolidating knowledge. Remarkably, it also enables the model to learn under high levels of label noise, which is ubiquitous in real-world data streams.

****

Auditing Fairness Online through Interactive Refinement

Pranav Maneriker,Codi Jay Burley,srinivasan parthasarathy

Machine learning algorithms are increasingly being deployed for high-stakes scenarios. A sizeable proportion of currently deployed models make their decisions in a black box manner. Such decision-making procedures are susceptible to intrinsic biases, which has led to a call for accountability in deployed decision systems. In this work, we focus on user-specified accountability of decision-making processes of black box systems. Previous work has formulated this problem as run time fairness monitoring over decision functions. However, formulating appropriate specifications for situation-appropriate fairness metrics is challenging. We construct AVOIR, an automated inference-based optimization system that improves bounds for and generalizes prior work across a wide range of fairness metrics. AVOIR offers an interactive and iterative process for exploring fairness violations aligned with governance and regulatory requirements. Our bounds improve over previous probabilistic guarantees for such fairness grammars in online settings. We also construct a novel visualization mechanism that can be used to investigate the context of reported fairness violations and guide users towards meaningful and compliant fairness specifications. We then conduct case studies with fairness metrics on three different datasets and demonstrate how the visualization and improved optimization can detect fairness violations more efficiently and ameliorate the issues with faulty fairness metric design.

****

REM: Routing Entropy Minimization for Capsule Networks

Riccardo Renzulli,Enzo Tartaglione,Marco Grangetto

Capsule Networks aim to build an interpretable and biologically-inspired neural network model. One of their main innovations relies on the routing mechanism which extracts a parse tree: its main purpose is to explicitly build relationships between capsules.

However, their true potential has not surfaced yet: these relationships are extremely heterogeneous and difficult to understand, as the intra-class extracted parse trees are very different from each other. A school of thoughts, giving-up on this side, propose less interpretable versions of Capsule Networks without routing.

This paper proposes REM, a technique which minimizes the entropy of the parse tree-like structure. We accomplish this by driving the model parameters distribution towards low entropy configurations, using a pruning mechanism as a proxy.

Thanks to REM, we generate a significantly lower number of parse trees, with essentially no performance loss, showing also that Capsule Networks build stronger and more stable relationships between capsules.

****

Don’t forget the nullspace! Nullspace occupancy as a mechanism for out of distribution failure

Daksh Idnani,Vivek Madan,Naman Goyal,David J. Schwab,Shanmukha Ramakrishna Vedantam

Out of distribution (OoD) generalization has received considerable interest in r ecent years. In this work, we identify a particular failure mode of OoD generalization for discriminative classifiers that is based on test data (from a new domain) lying in the nullspace of features learnt from source data. We demonstrate the existence of this failure mode across multiple networks trained across RotatedMNIST, PACS, TerraIncognita, DomainNet and ImageNet-R datasets. We then study different choices for characterizing the feature space and show that projecting intermediate representations onto the span of directions that obtain maximum training accuracy provides consistent improvements in OoD performance. Finally, we show that such nullspace behavior also provides an insight into neural networks trained on poisoned data. We hope our work galvanizes interest in the relationship between the nullspace occupancy failure mode and generalization.

****

Variational Classification

Shehzaad Zuzar Dhuliawala,Mrinmaya Sachan,Carl Allen

Classification tasks, ubiquitous across machine learning, are commonly tackled by a suitably designed neural network with a softmax output layer, mapping each data point to a categorical distribution over class labels.
We extend this familiar model from a latent variable perspective to variational classification (VC), analogous to how the variational auto-encoder relates to its deterministic counterpart. We derive a training objective based on the ELBO together with an \textit{adversarial} approach for optimising it.

Within this framework, we identify design choices made implicitly in off-the-shelf softmax functions and can instead include domain-specific assumptions, such as class-conditional latent priors. We demonstrate benefits of the VC model in image classification. We show on several standard datasets, that treating inputs to the softmax layer as latent variables under a mixture of Gaussians prior, improves several desirable aspects of a classifier, such as prediction accuracy, calibration, out-of-domain calibration and adversarial robustness.

****

ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond

Xiaojun Guo,Yifei Wang,Tianqi Du,Yisen Wang

Oversmoothing is a common phenomenon in a wide range of Graph Neural Networks (GNNs) and Transformers, where performance degenerates as the layer goes deeper. Instead of characterizing oversmoothing from the view of complete collapse in which representations converge to a single point, we dive into a more general perspective dimensional collapse in which representations lie in a narrow cone. Accordingly, inspired by the power of contrastive learning in preventing dimensional collapse, we propose a novel normalization layer ContraNorm. Intuitively, ContraNorm implicitly shatters representations in the embedding space, leading to a more uniform distribution and slighter dimensional collapse. On the theoretical analysis, we prove that ContraNorm can alleviate both complete collapse and dimensional collapse under some conditions. Our proposed normalization layer can be easily inserted into GNNs and Transformers with negligible parameter overhead. Experiments on various real-world datasets verify the effectiveness of our method.

****

Accelerated Single-Call Methods for Constrained Min-Max Optimization

Yang Cai,Weiqiang Zheng

We study first-order methods for constrained min-max optimization. Existing methods either require two gradient calls or two projections in each iteration, which may be costly in some applications. In this paper, we first show that a variant of the \emph{Optimistic Gradient (OG)} method, a \emph{single-call single-projection} algorithm, has $O(\frac{1}{\sqrt{T}})$ best-iterate convergence rate for inclusion problems with operators that satisfy the weak Minty variation inequality (MVI). Our second result is the first single-call single-projection algorithm -- the \emph{Accelerated Reflected Gradient (ARG)} method that achieves the \emph{optimal $O(\frac{1}{T})$} last-iterate convergence rate for inclusion problems that satisfy negative comonotonicity. Both the weak MVI and negative comonotonicity are well-studied assumptions and capture a rich set of non-convex non-concave min-max optimization problems. Finally, we show that the \emph{Reflected G radient (RG)} method, another \emph{single-call single-projection} algorithm, has $O(\frac{1}{\sqrt{T}})$ last-iterate convergence rate for constrained convex-concave min-max optimization, answering an open problem of [Hsieh et al., 2019]. Our convergence rates hold for standard measures such as the tangent residual and the natural residual.

****

Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes
Eoin M. Kenny,Mycal Tucker,Julie Shah
Despite recent success of deep learning models in research settings, their application in sensitive domains remains limited because of their opaque decision-making processes. Taking to this challenge, people have proposed various eXplainable AI (XAI) techniques designed to calibrate trust and understandability of black-box models, with the vast majority of work focused on supervised learning. Here, we focus on making an "interpretable-by-design" deep reinforcement learning agent which is forced to use human-friendly prototypes in its decisions, thus making its reasoning process clear. Our proposed method, dubbed Prototype-Wrapper Network (PW-Net), wraps around any neural agent backbone, and results indicate that it does not worsen performance relative to black-box models. Most importantly, we found in a user study that PW-Nets supported better trust calibration and task performance relative to standard interpretability approaches and black-boxes.

****

Distributed Extra-gradient with Optimal Complexity and Communication Guarantees
Ali Ramezani-Kebrya,Kimon Antonakopoulos,Igor Krawczuk,Justin Deschenaux,Volkan Cevher
We consider monotone variational inequality (VI) problems in multi-GPU settings where multiple processors/workers/clients have access to local stochastic dual vectors. This setting includes a broad range of important problems from distributed convex minimization to min-max and games. Extra-gradient, which is a de facto algorithm for monotone VI problems, has not been designed to be communication-efficient. To this end, we propose a quantized generalized extra-gradient (Q-GenX), which is an unbiased and adaptive compression method tailored to solve VIs. We provide an adaptive step-size rule, which adapts to the respective noise profiles at hand and achieve a fast rate of ${\cal O}(1/T)$ under relative noise, and an order-optimal ${\cal O}(1/\sqrt{T})$ under absolute noise and show distributed training accelerates convergence. Finally, we validate our theoretical results by providing real-world experiments and training generative adversarial networks on multiple GPUs.

****

'I pick you choose': Joint human-algorithm decision making in multi-armed bandits
Kate Donahue,Sreenivas Gollapudi,Kostas Kollias
Online learning in multi-armed bandits has been a rich area of research for decades, resulting in numerous \enquote{no-regret} algorithms that efficiently learn the arm with highest expected reward. However, in many settings the final decision of which arm to pull isn't under the control of the algorithm itself. For example, a driving app typically suggests a subset of routes (arms) to the driver, who ultimately makes the final choice about which to select. Typically, the human also wishes to learn the optimal arm based on historical reward information, but decides which arm to pull based on a potentially different objective function, such as being more or less myopic about exploiting near-term rewards. In this paper, we show when this joint human-algorithm system can achieve good performance. Specifically, we explore multiple possible frameworks for human objectives and give theoretical regret bounds for regret. Finally, we include experimental results exploring how regret varies with the human decision-maker's objective, as well as the number of arms presented.

****

What Matters In The Structured Pruning of Generative Language Models?
Michael Santacroce,Zixin Wen,yelong shen,Weizhu Chen,Yuanzhi Li

Auto-regressive large language models such as GPT-3 require enormous computational resources to use, leading to huge financial cost and environmental impact. Structured pruning methods traditionally reduce resource usage, however, their application to and efficacy for generative language models is heavily under-explored. We analyze the effects of magnitude, random, and movement (Lagunas et al., 2021) pruning on MLP layers in GPT-like models. We find that movement can under-perform for these models while random pruning nearly matches the best methods. By examining neuron-level redundancy measures, we discover that movement does not select neurons based on how unique they are compared to other neurons, leaving behind excess redundancy. In view of this, we introduce Globally Unique Movement (GUM) to select neurons based on both uniqueness and sensitivity. We then discuss the roles of our techniques on different redundancy metrics through careful comparisons and ablations.
****

MaxMin-Novelty: Maximizing Novelty via Minimizing the State-Action Values in Deep Reinforcement Learning
Ezgi Korkmaz
Reinforcement learning research has achieved high acceleration in its progress starting from the initial installation of deep neural networks as function approximators to learn policies that make sequential decisions in high-dimensional state representation MDPs. While several consecutive barriers have been broken in deep reinforcement learning research (i.e. learning from high-dimensional states, learning purely via self-play), several others still stand. On this line, the question of how to explore in high-dimensional complex MDPs is a well-understudied and ongoing open problem. To address this, in our paper we propose a unique exploration technique based on maximization of novelty via minimization of the state-action value function (MaxMin Novelty). Our method is theoretically well motivated, and comes with zero computational cost while leading to significant sample efficiency gains in deep reinforcement learning training. We conduct extensive experiments in the Arcade Learning Environment with high-dimensional state representation MDPs. We show that our technique improves the human normalized median scores of Arcade Learning Environment by 248% in the low-data regime.
****

Complete Likelihood Objective for Latent Variable Models
Mikhail Arkhipov,Maria Vikhreva
In this work, we propose an alternative to the Marginal Likelihood (MaL) objective for training latent variable models, Complete Latent Likelihood (CoLLike). We analyze the objectives from the perspective of matching joint distributions. We show that MaL corresponds to a particular KL divergence between some target \emph{joint} distribution and the model joint. Furthermore, the properties of the target joint explain such major malfunctions of MaL as uninformative latents (posterior collapse) and high deviation of the aggregated posterior from the prior. In CoLLike approach, we use a sample from the prior to construct a family of target joint distributions, which properties prevent these drawbacks. We utilize the complete likelihood both to choose the target from this family and to learn the model. We confirm our analysis by experiments with expressive low-dimensional latent variable models, which also indicate that it is possible to achieve high accuracy unsupervised classification using CoLLike objective.
****

The Surprising Effectiveness of Equivariant Models in Domains with Latent Symmetry
Dian Wang,Jung Yeon Park,Neel Sortur,Lawson L.S. Wong,Robin Walters,Robert Platt
Extensive work has demonstrated that equivariant neural networks can significantly improve sample efficiency and generalization by enforcing an inductive bias in the network architecture. These applications typically assume that the domain symmetry is fully described by explicit transformations of the model inputs and outputs. However, many real-life applications contain only latent or partial symmetries which cannot be easily described by simple transformations of the input. In these cases, it is necessary to learn symmetry in the environment instead of imposing it mathematically on the network architecture. We discover, surprising

ly, that imposing equivariance constraints that do not exactly match the domain symmetry is very helpful in learning the true symmetry in the environment. We differentiate between extrinsic and incorrect symmetry constraints and show that while imposing incorrect symmetry can impede the model's performance, imposing extrinsic symmetry can actually improve performance. We demonstrate that an equivariant model can significantly outperform non-equivariant methods on domains with latent symmetries both in supervised learning and in reinforcement learning for robotic manipulation and control problems.

****

Performance Bounds for Model and Policy Transfer in Hidden-parameter MDPs

Haotian Fu,Jiayu Yao,Omer Gottesman,Finale Doshi-Velez,George Konidaris

In the Hidden-Parameter MDP (HiP-MDP) framework, a family of reinforcement learning tasks is generated by varying hidden parameters specifying the dynamics and reward function for each individual task. HiP-MDP is a natural model for families of tasks in which meta- and lifelong-reinforcement learning approaches can succeed. Given a learned context encoder that infers the hidden parameters from previous experience, most existing algorithms fall into two categories: $\textit{model transfer}$ and $\textit{policy transfer}$, depending on which function the hidden parameters are used to parameterize. We characterize the robustness of model and policy transfer algorithms with respect to hidden parameter estimation error. We first show that the value function of HiP-MDPs is Lipschitz continuous under certain conditions. We then derive regret bounds for both settings through the lens of Lipschitz continuity. Finally, we empirically corroborate our theoretical analysis by experimentally varying the hyper-parameters governing the Lipschitz constants of two continuous control problems; the resulting performance is consistent with our predictions.

****

Parallel Q-Learning: Scaling Off-policy Reinforcement Learning

Zechu Li,Tao Chen,Zhang-Wei Hong,Anurag Ajay,Pulkit Agrawal

Reinforcement learning algorithms typically require tons of training data, resulting in long training time, especially on challenging tasks. With the recent advance in GPU-based simulation, such as Isaac Gym, data collection speed has been improved thousands of times on a commodity GPU. Most prior works have been using on-policy methods such as PPO to train policies in Isaac Gym due to its simpleness and effectiveness in scaling up. Off-policy methods are usually more sample-efficient but more challenging to be scaled up, resulting in a much longer wall-clock training time in practice. In this work, we presented a novel parallel Q-learning framework that not only gains better sample efficiency but also reduces the training wall-clock time compared to PPO. Different from prior works on distributed off-policy learning, such as Apex, our framework is designed specifically for massively parallel GPU-based simulation and optimized to work on a single workstation. We demonstrate the capability of scaling up Q learning methods to tens of thousands of parallel environments. We also investigate various factors that can affect the policy learning training speed, including the number of parallel environments, exploration schemes, batch size, GPU models, etc.

****

Emergence of shared sensory-motor graphical language from visual input

Yoann Lemesle,Tristan Karch,Clément Moulin-Frier,Romain Laroche,Pierre-Yves Oudeyer

The framework of Language Games studies the emergence of languages in populations of agents. Recent contributions relying on deep learning methods focused on agents communicating via an idealized communication channel, where utterances produced by a speaker are directly perceived by a listener. This comes in contrast with human communication, which instead relies on a sensory-motor channel, where motor commands produced by the speaker (e.g. vocal or gestural articulators) result in sensory effects perceived by the listener (e.g. audio or visual). Here, we investigate if agents can evolve a shared language when they are equipped with a continuous sensory-motor system to produce and perceive signs, e.g. drawings. To this end, we introduce the Graphical Referential Game (GREG) where a speaker must produce a graphical utterance to name a visual referent object consisting

of combinations of MNIST digits while a listener has to select the corresponding object among distractor referents, given the produced message. The utterances are drawing images produced using dynamical motor primitives combined with a sketching library. To tackle GREG we present CURVES: a multimodal contrastive deep learning mechanism that represents the energy (alignment) between named referents and utterances generated through gradient ascent on the learned energy landscape. We, then, present a set of experiments and metrics based on a systematic compositional dataset to evaluate the resulting language. We show that our method allows the emergence of a shared, graphical language with compositional properties
****

Composing Task Knowledge With Modular Successor Feature Approximators
Wilka Torrico Carvalho,Angelos Filos,Richard Lewis,Honglak Lee,Satinder Singh
Recently, the Successor Features and Generalized Policy Improvement (SF&GPI) framework has been proposed as a method for learning, composing and transferring predictive knowledge and behavior. SF&GPI works by having an agent learn predictive representations (SFs) that can be combined for transfer to new tasks with GPI. However, to be effective this approach requires state features that are useful to predict, and these state-features are typically hand-designed. In this work, we present a novel neural network architecture, “Modular Successor Feature Approximators” (MSFA), where modules both discover what is useful to predict, and learn their own predictive representations. We show that MSFA is able to better generalize compared to baseline architectures for learning SFs and a modular network that discovers factored state representations.

****

DexDeform: Dexterous Deformable Object Manipulation with Human Demonstrations and Differentiable Physics
Sizhe Li,Zhiao Huang,Tao Chen,Tao Du,Hao Su,Joshua B. Tenenbaum,Chuang Gan
In this work, we aim to learn dexterous manipulation of deformable objects using multi-fingered hands. Reinforcement learning approaches for dexterous rigid object manipulation would struggle in this setting due to the complexity of physics interaction with deformable objects. At the same time, previous trajectory optimization approaches with differentiable physics for deformable manipulation would suffer from local optima caused by the explosion of contact modes from hand-object interactions. To address these challenges, we propose DexDeform, a principled framework that abstracts dexterous manipulation skills from human demonstration, and refines the learned skills with differentiable physics. Concretely, we first collect a small set of human demonstrations using teleoperation. And we then train a skill model using demonstrations for planning over action abstractions in imagination. To explore the goal space, we further apply augmentations to the existing deformable shapes in demonstrations and use a gradient optimizer to refine the actions planned by the skill model. Finally, we adopt the refined trajectories as new demonstrations for finetuning the skill model. To evaluate the effectiveness of our approach, we introduce a suite of six challenging dexterous deformable object manipulation tasks. Compared with baselines, DexDeform is able to better explore and generalize across novel goals unseen in the initial human demonstrations. Additional materials can be found at our project website: https://sites.google.com/view/dexdeform.
****

NAG-GS: semi-implicit, accelerated and robust stochastic optimizer.
Valentin Leplat,Daniil Merkulov,Aleksandr Katrutsa,Daniel Bershatsky,Ivan Oseledets
Classical machine learning models such as deep neural networks are usually trained by using Stochastic Gradient Descent-based (SGD) algorithms. The classical SGD can be interpreted as a discretization of the stochastic gradient flow. In this paper we propose a novel, robust and accelerated stochastic optimizer that relies on two key elements: (1) an accelerated Nesterov-like Stochastic Differential Equation (SDE) and (2) its semi-implicit Gauss-Seidel type discretization. The convergence and stability of the obtained method, referred to as NAG-GS, are fi

rst studied extensively in the case of the minimization of a quadratic function. This analysis allows us to come up with an optimal step size (or learning rate) in terms of rate of convergence while ensuring the stability of NAG-GS. This is achieved by the careful analysis of the spectral radius of the iteration matrix and the covariance matrix at stationarity with respect to all hyperparameters of our method. We show that NAG-GS is competitive with state-of-the-art methods such as momentum SGD with weight decay and AdamW for the training of machine learning models such as the logistic regression model, the residual networks models on standard computer vision datasets, and Transformers in the frame of the GLUE benchmark.

****

Loop Unrolled Shallow Equilibrium Regularizer (LUSER) - A Memory-Efficient Inverse Problem Solver

Peimeng Guan,Jihui Jin,Justin Romberg,Mark A. Davenport

In inverse problems we aim to reconstruct some underlying signal of interest from potentially corrupted and often ill-posed measurements. Classical optimization-based techniques proceed by optimizing a data consistency metric together with a regularizer. Current state-of-the-art machine learning approaches draw inspiration from such techniques by unrolling the iterative updates for an optimization-based solver and then learning a regularizer from data. This \emph{loop unrolling} (LU) method has shown tremendous success, but often requires a deep model for the best performance leading to high memory costs during training. Thus, to address the balance between computation cost and network expressiveness, we propose an LU algorithm with shallow equilibrium regularizers (LUSER). These implicit models are as expressive as deeper convolutional networks, but far more memory efficient during training. The proposed method is evaluated on image deblurring, computed tomography (CT), as well as single-coil Magnetic Resonance Imaging (MRI) tasks and shows similar, or even better, performance while requiring up to $8\times$ less computational resources during training when compared against a more typical LU architecture with feedforward convolutional regularizers.

****

Robust Universal Adversarial Perturbations

Changming Xu,Gagandeep Singh

Universal Adversarial Perturbations (UAPs) are imperceptible, image-agnostic vectors that cause deep neural networks (DNNs) to misclassify inputs from a data distribution with high probability. In practical attack scenarios, adversarial perturbations may undergo transformations such as changes in pixel intensity, rotation, etc. while being added to DNN inputs. Existing methods do not create UAPs robust to these real-world transformations, thereby limiting their applicability in attack scenarios. In this work, we introduce and formulate robust UAPs. We build an iterative algorithm using probabilistic robustness bounds and transformations generated by composing arbitrary sub-differentiable transformation functions to construct such robust UAPs. We perform an extensive evaluation on the popular CIFAR-10 and ILSVRC 2012 datasets measuring our UAPs' robustness under a wide range common, real-world transformations such as rotation, contrast changes, etc. Our results show that our method can generate UAPs up to 23% more robust than existing state-of-the-art baselines.

****

Understanding the Complexity Gains of Contextual Multi-task RL with Curricula

Qiyang Li,Yuexiang Zhai,Yi Ma,Sergey Levine

Reinforcement learning (RL) problems can be challenging without well-shaped rewards. Prior work on provably efficient RL methods generally proposes to address this issue with dedicated exploration strategies, such as novelty-based bonuses. However, another way to tackle this challenge is to reformulate it as a multi-task RL problem, where the task space contains not only the challenging task of interest but also easier tasks that implicitly function as a curriculum. Such a reformulation opens up the possibility of running existing multi-task RL methods as a more efficient alternative to solving a single challenging task from scratch. In this work, we provide a theoretical framework that reformulates a single-task RL problem as a multi-task RL problem defined by a curriculum. Under mild reg ularity conditions on the curriculum, we show that sequentially solving each task in the multi-task RL problem is more computationally efficient than solving the original single-task problem, without any explicit exploration bonuses or other exploration strategies. We also show that our theoretical insights can be translated into an effective practical learning algorithm that can accelerate curriculum learning on simulated robotic goal-reaching tasks.

****

The Lie Derivative for Measuring Learned Equivariance

Nate Gruver,Marc Anton Finzi,Micah Goldblum,Andrew Gordon Wilson

Equivariance guarantees that a model's predictions capture key symmetries in data. When an image is translated or rotated, an equivariant model's representation of that image will translate or rotate accordingly. The success of convolutional neural networks has historically been tied to translation equivariance directly encoded in their architecture. The rising success of vision transformers, which have no explicit architectural bias towards equivariance, challenges this narrative and suggests that augmentations and training data might also play a significant role in their performance. In order to better understand the role of equivariance in recent vision models, we apply the Lie derivative, a method for measuring equivariance with strong mathematical foundations and minimal hyperparameters. Using the Lie derivative, we study the equivariance properties of hundreds of pretrained models, spanning CNNs, transformers, and Mixer architectures. The scale of our analysis allows us to separate the impact of architecture from other factors like model size or training method. Surprisingly, we find that many violations of equivariance can be linked to spatial aliasing in ubiquitous network layers, such as pointwise non-linearities, and that as models get larger and more accurate they tend to display more equivariance, regardless of architecture. For example, transformers can be more equivariant than convolutional neural networks after training.

****

Effective passive membership inference attacks in federated learning against overparameterized models

Jiacheng Li,Ninghui Li,Bruno Ribeiro

This work considers the challenge of performing membership inference attacks in a federated learning setting ---for image classification--- where an adversary can only observe the communication between the central node and a single client (a passive white-box attack). Passive attacks are one of the hardest-to-detect attacks, since they can be performed without modifying how the behavior of the central server or its clients, and assumes *no access to private data instances*. The key insight of our method is empirically observing that, near parameters that generalize well in test, the gradient of large overparameterized neural network models statistically behave like high-dimensional independent isotropic random vectors. Using this insight, we devise two attacks that are often little impacted by existing and proposed defenses. Finally, we validated the hypothesis that our attack depends on the overparametrization by showing that increasing the level of overparametrization (without changing the neural network architecture) positively correlates with our attack effectiveness.

****

Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning

Sheng Zhang,Hao Cheng,Jianfeng Gao,Hoifung Poon

We present a bi-encoder framework for named entity recognition (NER), which applies contrastive learning to map candidate text spans and entity types into the same vector representation space. Prior work predominantly approaches NER as sequence labeling or span classification. We instead frame NER as a representation learning problem that maximizes the similarity between the vector representations of an entity mention and its type. This makes it easy to handle nested and flat NER alike, and can better leverage noisy self-supervision signals. A major challenge to this bi-encoder formulation for NER lies in separating non-entity spans from entity mentions. Instead of explicitly labeling all non-entity spans as the same class $\texttt{Outside}$ ($\texttt{O}$) as in most prior methods, we introduce a novel dynamic thresholding loss, learned in conjunction with the standar d contrastive loss. Experiments show that our method performs well in both supervised and distantly supervised settings, for nested and flat NER alike, establishing new state of the art across standard datasets in the general domain (e.g., ACE2004, ACE2005, CoNLL2003) and high-value verticals such as biomedicine (e.g., GENIA, NCBI, BC5CDR, JNLPBA). We release the code at https://github.com/microsoft/binder.

****

Handling Covariate Shifts in Federated Learning with Generalization Guarantees
Ali Ramezani-Kebrya,Fanghui Liu,Thomas Pethick,Grigorios Chrysos,Volkan Cevher
Covariate shift across clients is a major challenge for federated learning (FL). This work studies the generalization properties of FL under intra-client and inter-client covariate shifts. To this end, we propose Federated Importance-weighteD Empirical risk Minimization (FIDEM) to optimize a global FL model, along with new variants of density ratio matching methods, aiming to handle covariate shifts. These methods trade off some level of privacy for improving the overall generalization performance. We theoretically show that FIDEM achieves smaller generalization error than classical empirical risk minimization under some certain settings. Experimental results demonstrate the superiority of FIDEM over federated averaging (McMahan et al., 2017) and other baselines, which would open the door to study FL under distribution shifts more systematically.

****

Agree to Disagree: Diversity through Disagreement for Better Transferability
Matteo Pagliardini,Martin Jaggi,François Fleuret,Sai Praneeth Karimireddy
Gradient-based learning algorithms have an implicit \emph{simplicity bias} which in effect can limit the diversity of predictors being sampled by the learning procedure. This behavior can hinder the transferability of trained models by (i) favoring the learning of simpler but spurious features --- present in the training data but absent from the test data --- and (ii) by only leveraging a small subset of predictive features. Such an effect is especially magnified when the test distribution does not exactly match the train distribution---referred to as the Out of Distribution (OOD) generalization problem. However, given only the training data, it is not always possible to apriori assess if a given feature is spurious or transferable. Instead, we advocate for learning an ensemble of models which capture a diverse set of predictive features. Towards this, we propose a new algorithm D-BAT (Diversity-By-disAgreement Training), which enforces agreement among the models on the training data, but disagreement on the OOD data. We show how D-BAT naturally emerges from the notion of generalized discrepancy, as well as demonstrate in multiple experiments how the proposed method can mitigate shortcut-learning, enhance uncertainty and OOD detection, as well as improve transferability.

****

Expected Probabilistic Hierarchies
Marcel Kollovieh,Bertrand Charpentier,Daniel Zügner,Stephan Günnemann
Hierarchical clustering has usually been addressed by discrete optimization using heuristics or continuous optimization of relaxed scores for hierarchies. In this work, we propose to optimize expected scores under a probabilistic model over hierarchies. (1) We show theoretically that the global optimum of the expected Dasgupta cost and Tree-Sampling divergence (TSD), two unsupervised metrics for hierarchical clustering scores, are equal to the optimum of their discrete counterparts contrary to some relaxed scores. (2) We propose Expected Probabilistic Hierarchies (EPH), a probabilistic model to learn hierarchies in data by optimizing expected scores. EPH uses differentiable hierarchy sampling enabling end-to-end gradient-descent based optimizations, and an unbiased subgraph sampling approach to scale to large datasets. (3) We evaluate EPH on synthetic and real-world datasets including vector and graph datasets. EPH outperforms all other approaches on quantitative results and provides meaningful hierarchies in qualitative evaluations.

****

Taking a Step Back with KCal: Multi-Class Kernel-Based Calibration for Deep Neur al Networks

Zhen Lin,Shubhendu Trivedi,Jimeng Sun

Deep neural network (DNN) classifiers are often overconfident, producing miscalibrated class probabilities. In high-risk applications like healthcare, practitioners require fully calibrated probability predictions for decision-making. That is, conditioned on the prediction vector, every class’ probability should be close to the predicted value. Most existing calibration methods either lack theoretical guarantees for producing calibrated outputs, reduce classification accuracy in the process, or only calibrate the predicted class. This paper proposes a new Kernel-based calibration method called KCal. Unlike existing calibration procedures, KCal does not operate directly on the logits or softmax outputs of the DNN. Instead, KCal learns a metric space on the penultimate-layer latent embedding and generates predictions using kernel density estimates on a calibration set. We first analyze KCal theoretically, showing that it enjoys a provable full calibration guarantee. Then, through extensive experiments across a variety of datasets, we show that KCal consistently outperforms baselines as measured by the calibration error and by proper scoring rules like the Brier Score.

****

A distinct unsupervised reference model from the environment helps continual learning

Seyyed AmirHossein Ameli Kalkhoran,Mohammadamin Banayeeanzade,Mahdi Samiei,Mahdieh Soleymani Baghshah

The existing continual learning methods are mainly focused on fully-supervised scenarios and are still not able to take advantage of unlabeled data available in the environment. Some recent works tried to investigate semi-supervised continual learning (SSCL) settings in which the unlabeled data are available, but it is only from the same distribution as the labeled data. This assumption is still not general enough for real-world applications and restricts the utilization of unsupervised data. In this work, we introduce Open-Set Semi-Supervised Continual Learning (OSSCL), a more realistic semi-supervised continual learning setting in which out-of-distribution (OoD) unlabeled samples in the environment are assumed to coexist with the in-distribution ones. Under this configuration, we present a model with two distinct parts: (i) the reference network captures general-purpose and task-agnostic knowledge in the environment by using a broad spectrum of unlabeled samples, (ii) the learner network is designed to learn task-specific representations by exploiting supervised samples. The reference model both provides a pivotal representation space and also segregates unlabeled data to exploit them more efficiently. By performing a diverse range of experiments, we show the superior performance of our model compared with other competitors and prove the effectiveness of each component of the proposed model.

****

The Crossword Puzzle: Simplifying Deep Neural Network Pruning with Fabulous Coordinates

Yicun Duan,Wangkai Jin,Xiangjun Peng

Pruning is a promising technique to shrink the size of Deep Neural Network models with only negligible accuracy overheads. Recent efforts rely on experience-derived metric to guide pruning procedure, which heavily saddles with the effective generalization of pruning methods. We propose The Cross Puzzle, a new method to simplify this procedure by automatically deriving pruning metrics. The key insight behind our method is that: \textit{For Deep Neural Network Models, a Pruning-friendly Distribution of model's weights can be obtained, given a proper Coordinate}. We experimentally confirm the above insight, and denote the new Coordinate as the Fabulous Coordinates. Our quantitative evaluation results show that: the Crossword Puzzle can find a simple yet effective metric, which outperforms the state-of-the-art pruning methods by delivering no accuracy degradation on ResNet-56 (CIFAR-10)/-101 (ImageNet), while the pruning rate is raised to 70\%/50\% for the respective models.

****

Learning the Visualness of Text Using Large Vision-Language Models

Gaurav Verma,Ryan A. Rossi,Christopher Tensmeyer,Jiuxiang Gu,Ani Nenkova

Visual text evokes an image in a person's mind, while non-visual text fails to do so. A method to automatically detect visual text will unlock the ability to augment text with relevant images, as neural text-to-image generation and retrieval models operate on the implicit assumption that the input text is visual in nature. We curate a dataset of 3,620 English sentences and their visualness scores provided by multiple human annotators. Additionally, we use documents that contain text and visual assets to create a distantly supervised corpus of document text and associated images. We also propose a fine-tuning strategy that adapts large vision-language models like CLIP that assume a one-to-one correspondence between text and image to the task of scoring text visualness from text input alone. Our strategy involves modifying the model's contrastive learning objective to map text identified as non-visual to a common NULL image while matching visual text to their corresponding images in the document. We evaluate the proposed approach on its ability to (i) classify visual and non-visual text accurately, and (ii) attend over words that are identified as visual in psycholinguistic studies. Empirical evaluation indicates that our approach performs better than several heuristics and baseline models for the proposed task. Furthermore, to highlight the importance of modeling the visualness of text, we conduct qualitative analyses of text-to-image generation systems like DALL-E. We release the curated dataset and code.

****

SemPPL: Predicting Pseudo-Labels for Better Contrastive Representations

Matko Bošnjak,Pierre Harvey Richemond,Nenad Tomasev,Florian Strub,Jacob C Walker,Felix Hill,Lars Holger Buesing,Razvan Pascanu,Charles Blundell,Jovana Mitrovic

Learning from large amounts of unsupervised data and a small amount of supervision is an important open problem in computer vision. We propose a new semi-supervised learning method, Semantic Positives via Pseudo-Labels (SEMPPL), that combines labelled and unlabelled data to learn informative representations. Our method extends self-supervised contrastive learning—where representations are shaped by distinguishing whether two samples represent the same underlying datum (positives) or not (negatives)—with a novel approach to selecting positives. To enrich the set of positives, we leverage the few existing ground-truth labels to predict the missing ones through a k-nearest neighbors classifier by using the learned embeddings of the labelled data. We thus extend the set of positives with datapoints having the same pseudo-label and call these semantic positives. We jointly learn the representation and predict bootstrapped pseudo-labels. This creates a reinforcing cycle. Strong initial representations enable better pseudo-label predictions which then improve the selection of semantic positives and lead to even better representations. SEMPPL outperforms competing semi-supervised methods setting new state-of-the-art performance of 76% and 68.5% top-1accuracy when using a ResNet-50 and training on 10% and 1% of labels on ImageNet, respectively. Furthermore, when using selective kernels, SEMPPL significantly outperforms previous state-of-the-art achieving 72.3% and 78.3% top-1accuracy on ImageNet with 1% and 10% labels, respectively, which improves absolute +7.8% and +6.2% over previous work. SEMPPL also exhibits state-of-the-art performance over larger ResNet models as well as strong robustness, out-of-distribution and transfer performance. We release the checkpoints and the evaluation code at https://github.com/deepmind/semppl.

****

Differentially Private Adaptive Optimization with Delayed Preconditioners

Tian Li,Manzil Zaheer,Ken Liu,Sashank J. Reddi,Hugh Brendan McMahan,Virginia Smith

Privacy costs may negate the benefits of using adaptive optimizers in differentially private model training. Prior works typically address this issue by using auxiliary information (e.g., public data) to boost the effectiveness of adaptive optimization. In this work, we explore techniques to estimate and efficiently adapt to gradient geometry in private adaptive optimization without auxiliary data. Motivated by the observation that adaptive methods can tolerate stale preconditioners, we propose differentially private adaptive training with delayed preconditioners (DP^2), a simple method that constructs delayed but less noisy precond itioners to better realize the benefits of adaptivity. Theoretically, we provide convergence guarantees for our method for both convex and non-convex problems, and analyze trade-offs between delay and privacy noise reduction. Empirically, we explore DP^2 across several real-world datasets, demonstrating that it can improve convergence speed by as much as 4× relative to non-adaptive baselines and match the performance of state-of-the-art optimization methods that require auxiliary data.

****

Towards a Mathematics Formalisation Assistant using Large Language Models

Ayush Agrawal,Siddhartha Gadgil,Navin Goyal,Ashvni Narayanan,Anand Tadipatri

Mathematics formalisation is the task of writing mathematics (i.e., definitions, theorem statements, proofs) in natural language, as found in books and papers, into a formal language that can then be checked for correctness by a program. It is a thriving activity today, however formalisation remains cumbersome. In this paper, we explore the abilities of a large language model (Codex) to help with formalisation in the Lean theorem prover. We find that with careful input-dependent prompt selection and postprocessing, Codex is able to formalise short mathematical statements at undergrad level with nearly 75\% accuracy for 120 theorem statements. For proofs quantitative analysis is infeasible and we undertake a detailed case study. We choose a diverse set of 13 theorems at undergrad level with proofs that fit in two-three paragraphs. We show that with a new prompting strategy Codex can formalise these proofs in natural language with at least one out of twelve Codex completion being easy to repair into a complete proof. This is surprising as essentially no aligned data exists for formalised mathematics, particularly for proofs. These results suggest that large language models are a promising avenue towards fully or partially automating formalisation.

****

FedLite: Improving Communication Efficiency in Federated Split Learning

Jianyu Wang,Hang Qi,Ankit Singh Rawat,Sashank J. Reddi,Sagar M. Waghmare,Felix Yu,Gauri Joshi

In classical federated learning, clients contribute to the overall training by communicating local updates for the underlying model on their private data to a coordinating server. However, updating and communicating the entire model becomes prohibitively expensive when resource-constrained clients collectively aim to train a large machine learning model. Split learning provides a natural solution in such a setting, where only a (small) part of the model is stored and trained on clients while the remaining (large) part of the model only stays at the servers. Unfortunately, the model partitioning employed in split learning significantly increases the communication cost compared to the classical federated learning algorithms. This paper addresses this issue by compressing the additional communication cost associated with split learning via a novel clustering algorithm and a gradient correction technique. An extensive empirical evaluation on standard image and text benchmarks shows that the proposed method can achieve up to 490x communication cost reduction with minimal drop in accuracy, and enables a desirable performance vs. communication trade-off.

****

Adversarial Policies Beat Professional-Level Go AIs

Tony Tong Wang,Adam Gleave,Nora Belrose,Tom Tseng,Michael D Dennis,Yawen Duan,Viktor Pogrebniak,Sergey Levine,Stuart Russell

We attack the state-of-the-art Go-playing AI system, KataGo, by training an adversarial policy that plays against a frozen KataGo victim. Our attack achieves a >99% win-rate against KataGo without search, and a >80% win-rate when KataGo uses enough search to be near-superhuman. To the best of our knowledge, this is the first successful end-to-end attack against a Go AI playing at the level of a top human professional. Notably, the adversary does not win by learning to play Go better than KataGo---in fact, the adversary is easily beaten by human amateurs. Instead, the adversary wins by tricking KataGo into ending the game prematurely at a point that is favorable to the adversary. Our results demonstrate that even professional-level AI systems may harbor surprising failure modes.

****

Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions
Ruben Villegas,Mohammad Babaeizadeh,Pieter-Jan Kindermans,Hernan Moraldo,Han Zhang,Mohammad Taghi Saffar,Santiago Castro,Julius Kunze,Dumitru Erhan

We present Phenaki, a model capable of realistic video synthesis given a sequence of textual prompts. Generating videos from text is particularly challenging due to the computational cost, limited quantities of high quality text-video data and variable length of videos. To address these issues, we introduce a new causal model for learning video representation which compresses the video to a small discrete tokens representation. This tokenizer is auto-regressive in time, which allows it to work with video representations of different length.

To generate video tokens from text we are using a bidirectional masked transformer conditioned on pre-computed text tokens. The generated video tokens are subsequently de-tokenized to create the actual video. To address data issues, we demonstrate how joint training on a large corpus of image-text pairs as well as a smaller number of video-text examples can result in generalization beyond what is available in the video datasets. Compared to the previous video generation methods, Phenaki can generate arbitrary long videos conditioned on a sequence of prompts (i.e. time variable text or story) in open domain. To the best of our knowledge, this is the first time a paper studies generating videos from time variable prompts.

****

Long Range Language Modeling via Gated State Spaces
Harsh Mehta,Ankit Gupta,Ashok Cutkosky,Behnam Neyshabur

State space models have shown to be effective at modeling long range dependencies, specially on sequence classification tasks. In this work we focus on autoregressive sequence modeling over English books, Github source code and ArXiv mathematics articles. Based on recent developments around the effectiveness of gated activation functions, we propose a new layer named \textit{Gated State Space} (GSS) and show that it trains significantly faster than the diagonal version of S4 (i.e. DSS) on TPUs, is fairly competitive with several well-tuned Transformer-based baselines and exhibits zero-shot generalization to longer inputs while being straightforward to implement. Finally, we show that leveraging self-attention to model local dependencies improves the performance of GSS even further.

****

On the (Non-)Robustness of Two-Layer Neural Networks in Different Learning Regimes
Elvis Dohmatob,Alberto Bietti

Neural networks are known to be highly sensitive to adversarial examples. These may arise due to different factors, such as random initialization, or spurious correlations in the learning problem. To better understand these factors, we provide a precise study of the adversarial robustness in different scenarios, from initialization to the end of training in different regimes, as well as intermediate scenarios where initialization still plays a role due to “lazy” training. We consider over-parameterized networks in high dimensions with quadratic targets and infinite samples. Our analysis allows us to identify new tradeoffs between approximation (as measured via test error) and robustness, whereby robustness can only get worse when test error improves, and vice versa. We also show how linearized lazy training
regimes can worsen robustness, due to improperly scaled random initialization. Our theoretical results are illustrated with numerical experiments.

****

Task-customized Masked Autoencoder via Mixture of Cluster-conditional Experts
Zhili LIU,Kai Chen,Jianhua Han,Lanqing HONG,Hang Xu,Zhenguo Li,James Kwok

Masked Autoencoder (MAE) is a prevailing self-supervised learning method that achieves promising results in model pre-training. However, when the various downstream tasks have data distributions different from the pre-training data, the semantically irrelevant pre-training information might result in negative transfer, impeding MAE’s scalability. To address this issue, we propose a novel MAE-based pre-training paradigm, Mixture of Cluster-conditional Experts (MoCE), which can be trained once but provides customized pre-training models for diverse downstr eam tasks. Different from the mixture of experts (MoE), our MoCE trains each expert only with semantically relevant images by using cluster-conditional gates. Thus, each downstream task can be allocated to its customized model pre-trained with data most similar to the downstream data. Experiments on a collection of 11 downstream tasks show that MoCE outperforms the vanilla MAE by 2.45\% on average. It also obtains new state-of-the-art self-supervised learning results on detection and segmentation.

****

A Deep Dive into Dataset Imbalance and Bias in Face Identification

Valeriia Cherepanova,Steven Reich,Samuel Dooley,Hossein Souri,John P Dickerson,Micah Goldblum,Tom Goldstein

As the deployment of automated face recognition (FR) systems proliferates, bias in these systems is not just an academic question, but a matter of public concern. Media portrayals often center imbalance as the main source of bias, i.e., that FR models perform worse on images of non-white people or women because these demographic groups are underrepresented in training data. Recent academic research paints a more nuanced picture of this relationship. However, previous studies of data imbalance in FR have focused exclusively on the face verification setting, while the face identification setting has been largely ignored, despite being deployed in sensitive applications such as law enforcement. This is an unfortunate omission, as 'imbalance' is a more complex matter in identification; imbalance may arise in not only the training data, but also the testing data, and furthermore may affect the proportion of identities belonging to each demographic group or the number of images belonging to each identity. In this work, we address this gap in the research by thoroughly exploring the effects of each kind of imbalance possible in face identification, and discuss other factors which may impact bias in this setting.

****

Causally Constrained Data Synthesis For Private Data Release

Varun Chandrasekaran,Darren Edge,Somesh Jha,Lukas Wutschitz,Amit Sharma,Cheng Zhang,Shruti Tople

Data privacy is critical in many decision-making contexts, such as healthcare and finance. A common mechanism is to create differentially private synthetic data using generative models. Such data generation reflects certain statistical properties of the original data, but often has an unacceptable privacy vs. utility trade-off. Since natural data inherently exhibits causal structure, we propose incorporating \emph{causal information} into the training process to favorably navigate the aforementioned trade-off. Under certain assumptions for linear gaussian models and a broader class of models, we theoretically prove that causally informed generative models provide better differential privacy guarantees than their non-causal counterparts. We evaluate our proposal using variational autoencoders, and demonstrate that the trade-off is mitigated through better utility for comparable privacy.

****

Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization

Jivat Neet Kaur,Emre Kiciman,Amit Sharma

Recent empirical studies on domain generalization (DG) have shown that DG algorithms that perform well on some distribution shifts fail on others, and no state-of-the-art DG algorithm performs consistently well on all shifts. Moreover, real-world data often has multiple distribution shifts over different attributes; hence we introduce multi-attribute distribution shift datasets and find that the accuracy of existing DG algorithms falls even further. To explain these results, we provide a formal characterization of generalization under multi-attribute shifts using a canonical causal graph. Based on the relationship between spurious attributes and the classification label, we obtain realizations of the canonical causal graph that characterize common distribution shifts and show that each shift entails different independence constraints over observed variables. As a result, we prove that any algorithm based on a single, fixed constraint cannot work well across all shifts, providing theoretical evidence for mixed empirical resul ts on DG algorithms. Based on this insight, we develop Causally Adaptive Constraint Minimization (CACM), an algorithm that uses knowledge about the data-generating process to adaptively identify and apply the correct independence constraints for regularization. Results on fully synthetic, MNIST, small NORB, and Waterbirds datasets, covering binary and multi-valued attributes and labels, show that adaptive dataset-dependent constraints lead to the highest accuracy on unseen domains whereas incorrect constraints fail to do so. Our results demonstrate the importance of modeling the causal relationships inherent in the data-generating process.

****

Bayes-MIL: A New Probabilistic Perspective on Attention-based Multiple Instance Learning for Whole Slide Images

Yufei CUI,Ziquan Liu,Xiangyu Liu,Xue Liu,Cong Wang,Tei-Wei Kuo,Chun Jason Xue,Antoni B. Chan

Multiple instance learning (MIL) is a popular weakly-supervised learning model on the whole slide image (WSI) for AI-assisted pathology diagnosis. The recent advance in attention-based MIL allows the model to find its region-of-interest (ROI) for interpretation by learning the attention weights for image patches of WSI slides. However, we empirically find that the interpretability of some related methods is either untrustworthy as the principle of MIL is violated or unsatisfactory as the high-attention regions are not consistent with experts' annotations. In this paper, we propose Bayes-MIL to address the problem from a probabilistic perspective. The induced patch-level uncertainty is proposed as a new measure of MIL interpretability, which outperforms previous methods in matching doctors annotations. We design a slide-dependent patch regularizer (SDPR) for the attention, imposing constraints derived from the MIL assumption, on the attention distribution. SDPR explicitly constrains the model to generate correct attention values. The spatial information is further encoded by an approximate convolutional conditional random field (CRF), for better interpretability. Experimental results show Bayes-MIL outperforms the related methods in patch-level and slide-level metrics and provides much better interpretable ROI on several large-scale WSI datasets.

****

Exploring Connections Between Memorization And Membership Inference

Jihye Choi,Varun Chandrasekaran,Shruti Tople,Somesh Jha

Membership inference (MI) allows privacy adversaries to query trained machine learning models to infer if a particular data sample was used in model training. Prior work has shown that the efficacy of MI is not the same for every sample in the training dataset; they broadly attribute this behavior to various data properties such as distributional difference. However, systematically analyzing the reasons for such disparate behavior has received little attention. In this work, we investigate the cause for such a discrepancy, and observe that the reason is more subtle and fundamental. We first provide empirical insight that an MI adversary is very successful with those samples that are highly $\textit{likely to be memorized}$, irrespective of whether the sample is from the same or a different distribution. Next, we provide a game-based formulation which lower-bounds the advantage of an adversary with the ability to determine if a sample is memorized or not, under certain assumptions made about the efficacy of the model on the memorized samples. Finally, based on our theoretical results, we present a practical instantiation of a highly effective MI attack on memorized samples.

****

Action Matching: A Variational Method for Learning Stochastic Dynamics from Samples

Kirill Neklyudov,Daniel Severo,Alireza Makhzani

Stochastic dynamics are ubiquitous in many fields of science, from the evolution of quantum systems in physics to diffusion-based models in machine learning. Existing methods such as score matching can be used to simulate these physical processes by assuming that the dynamics is a diffusion, which is not always the case. In this work, we propose a method called "Action Matching" that enables us to learn a much broader family of stochastic dynamics. Our method requires access only to samples from different time-steps, makes no explicit assumptions about the underlying dynamics, and can be applied even when samples are uncorrelated (i.e., are not part of a trajectory). Action Matching directly learns an underlying mechanism to move samples in time without modeling the distributions at each time-step. In this work, we showcase how Action Matching can be used for several computer vision tasks such as generative modeling, super-resolution, colorization, and inpainting; and further discuss potential applications in other areas of science.

****

Pre-train Graph Neural Networks for Brain Network Analysis

Yi Yang,Hejie Cui,Carl Yang

Human brains, controlling behaviors and cognition, are at the center of complex neurobiological systems. Recent studies in neuroscience and neuroimaging analysis have reached a consensus that interactions among brain regions of interest (ROIs) are driving factors for neural development and disorders. Graph neural networks as a powerful tool for analyzing graph-structured data are naturally applied to the analysis of brain networks. However, training of deep learning models including GNNs often requires a significant amount of labeled data. Due to the complicated data acquisition process and restrictions on data sharing, brain network datasets are still small compared to other domains (e.g., molecules, proteins). Moreover, real clinical tasks (e.g., mental disorder analysis) are often conducted on local datasets with even smaller scales and larger noises. To this end, we propose to leverage pre-training to capture the intrinsic brain network structures regardless of specific clinical outcomes. Specifically, we characterize the contributions in this work from two perspectives: (1) We design brain-network-oriented unsupervised pre-training techniques to utilize large-scale brain imaging studies without highly relevant task labels. (2) To facilitate effective knowledge transfer across studies with different ROI systems, we propose to develop a data-driven parcellation atlas mapping pipeline. The proposed pre-training techniques are validated with various GNN models. Extensive experiments demonstrate consistent improvement in performance as well as robustness.

****

Investigating Multi-task Pretraining and Generalization in Reinforcement Learning

Adrien Ali Taiga,Rishabh Agarwal,Jesse Farebrother,Aaron Courville,Marc G Bellemare

Deep reinforcement learning~(RL) has achieved remarkable successes in complex single-task settings. However, designing RL agents that can learn multiple tasks and leverage prior experience to quickly adapt to a related new task remains challenging. Despite previous attempts to improve on these areas, our understanding of multi-task training and generalization in RL remains limited. To fill this gap, we investigate the generalization capabilities of a popular actor-critic method, IMPALA. Specifically, we build on previous work that has advocated for the use of modes and difficulties of Atari 2600 games as a challenging benchmark for transfer learning in RL. We do so by pretraining an agent on multiple variants of the same Atari game before fine-tuning on the remaining never-before-seen variants. This protocol simplifies the multi-task pretraining phase by limiting negative interference between tasks and allows us to better understand the dynamics of multi-task training and generalization. We find that, given a fixed amount of pretraining data, agents trained with more variations are able to generalize better. Surprisingly, we also observe that this advantage can still be present after fine-tuning for 200M environment frames than when doing zero-shot transfer. This highlights the potential effect of a good learned representation. We also find that, even though small networks have remained popular to solve Atari 2600 games, increasing the capacity of the value and policy network is critical to achieve good performance as we increase the number of pretraining modes and difficulties. Overall, our findings emphasize key points that are essential for efficient multi-task training and generalization in reinforcement learning.

****

FIT: A Metric for Model Sensitivity

Ben Zandonati,Adrian Alan Pol,Maurizio Pierini,Olya Sirkin,Tal Kopetz

Model compression is vital to the deployment of deep learning on edge devices. Low precision representations, achieved via quantization of weights and activations, can reduce inference time and memory requirements. However, quantifying and predicting the response of a model to the changes associated with this procedure remains challenging. This response is non-linear and heterogeneous throughout the network. Understanding which groups of parameters and activations are more sensitive to quantization than others is a critical stage in maximizing efficiency. For this purpose, we propose FIT. Motivated by an information geometric perspective, FIT combines the Fisher information with a model of quantization. We find that FIT can estimate the final performance of a network without retraining. FIT effectively fuses contributions from both parameter and activation quantization into a single metric. Additionally, FIT is fast to compute when compared to existing methods, demonstrating favourable convergence properties. These properties are validated experimentally across hundreds of quantization configurations, with a focus on layer-wise mixed-precision quantization.

****

Transfer Learning with Deep Tabular Models

Roman Levin,Valeriia Cherepanova,Avi Schwarzschild,Arpit Bansal,C. Bayan Bruss,Tom Goldstein,Andrew Gordon Wilson,Micah Goldblum

Recent work on deep learning for tabular data demonstrates the strong performance of deep tabular models, often bridging the gap between gradient boosted decision trees and neural networks. Accuracy aside, a major advantage of neural models is that they are easily fine-tuned in new domains and learn reusable features. This property is often exploited in computer vision and natural language applications, where transfer learning is indispensable when task-specific training data is scarce. In this work, we explore the benefits that representation learning provides for knowledge transfer in the tabular domain. We conduct experiments in a realistic medical diagnosis test bed with limited amounts of downstream data and find that transfer learning with deep tabular models provides a definitive advantage over gradient boosted decision tree methods. We further compare the supervised and self-supervised pretraining strategies and provide practical advice on transfer learning with tabular models. Finally, we propose a pseudo-feature method for cases where the upstream and downstream feature sets differ, a tabular-specific problem widespread in real-world applications.

****

An Empirical Study on the Efficacy of Deep Active Learning Techniques

YU LI,Muxi Chen,Yannan Liu,Daojing He,Qiang Xu

Deep Active Learning (DAL) has been advocated as a promising method to reduce labeling costs in supervised learning. However, existing evaluations of DAL methods are based on different settings, and their results are controversial. To tackle this issue, this paper comprehensively evaluates 19 existing DAL methods in a uniform setting, including traditional fully-\underline{s}upervised \underline{a}ctive \underline{l}earning (SAL) strategies and emerging \underline{s}emi-\underline{s}upervised \underline{a}ctive \underline{l}earning (SSAL) techniques. We have several non-trivial findings. First, most SAL methods cannot achieve higher accuracy than random selection. Second, semi-supervised training brings significant performance improvement compared to pure SAL methods. Third, performing data selection in the SSAL setting can achieve a significant and consistent performance improvement, especially with abundant unlabeled data. Our findings produce the following guidance for practitioners: one should (i) apply SSAL as early as possible and (ii) collect more unlabeled data whenever possible, for better model performance. We will release our code upon acceptance.

****

CrAM: A Compression-Aware Minimizer

Alexandra Peste,Adrian Vladu,Eldar Kurtic,Christoph H Lampert,Dan Alistarh

Deep neural networks (DNNs) often have to be compressed, via pruning and/or quantization, before they can be deployed in practical settings. In this work we propose a new compression-aware minimizer dubbed CrAM that modifies the optimization step in a principled way, in order to produce models whose local loss behavior is stable under compression operations such as pruning. Thus, dense models trained via CrAM should be compressible post-training, in a single step, without significant accuracy loss. Experimental results on standard benchmarks, such as residual networks for ImageNet classification and BERT models for language modelling, show that CrAM produces dense models that can be more accurate than the standard SGD/Adam-based baselines, but which are stable under weight pruning: specifically, we can prune models in one-shot to 70-80% sparsity with almost no accuracy loss, and to 90% with reasonable (~ 1%) accuracy loss, which is competitive with gradual compression methods. Additionally, CrAM can produce sparse models which perform well for transfer learning, and it also works for semi-structured 2:4 pruning patterns supported by GPU hardware. The code for reproducing the results is available at: https://github.com/IST-DASLab/CrAM .

****

Using Language to Extend to Unseen Domains

Lisa Dunlap,Clara Mohri,Devin Guillory,Han Zhang,Trevor Darrell,Joseph E. Gonzalez,Aditi Raghunathan,Anna Rohrbach

It is expensive to collect training data for every possible domain that a vision model may encounter when deployed. We instead consider how simply $\textit{verbalizing}$ the training domain (e.g.``photos of birds'') as well as domains we want to extend to but do not have data for (e.g.``paintings of birds'') can improve robustness. Using a multimodal model with a joint image and language embedding space, our method $\textit{LADS}$ learns a transformation of the image embeddings from the source domain to each target domain, while preserving task relevant information. Without using any images from the target domain, we show that over the $\textit{extended}$ domain containing both source and target, $\textit{LADS}$ outperforms standard fine-tuning and ensemble approaches over a suite of 4 benchmarks targeting domain adaptation and dataset bias.

****

Can We Find Nash Equilibria at a Linear Rate in Markov Games?

Zhuoqing Song,Jason D. Lee,Zhuoran Yang

We study decentralized learning in two-player zero-sum discounted Markov games where the goal is to design a policy optimization algorithm for either agent satisfying two properties. First, the player does not need to know the policy of the opponent to update its policy. Second, when both players adopt the algorithm, their joint policy converges to a Nash equilibrium of the game. To this end, we construct a meta-algorithm, dubbed as $\texttt{Homotopy-PO}$, which provably finds a Nash equilibrium at a global linear rate. In particular, $\texttt{Homotopy-PO}$ interweaves two base algorithms $\texttt{Local-Fast}$ and $\texttt{Global-Slow}$ via homotopy continuation. $\texttt{Local-Fast}$ is an algorithm that enjoys local linear convergence while $\texttt{Global-Slow}$ is an algorithm that converges globally but at a slower sublinear rate. By switching between these two base algorithms, $\texttt{Global-Slow}$ essentially serves as a ``guide'' which identifies a benign neighborhood where $\texttt{Local-Fast}$ enjoys fast convergence. However, since the exact size of such a neighborhood is unknown, we apply a doubling trick to switch between these two base algorithms. The switching scheme is delicately designed so that the aggregated performance of the algorithm is driven by $\texttt{Local-Fast}$. Furthermore, we prove that $\texttt{Local-Fast}$ and $\texttt{Global-Slow}$ can both be instantiated by variants of optimistic gradient descent/ascent (OGDA) method, which is of independent interest.

****

Speeding up Policy Optimization with Vanishing Hypothesis and Variable Mini-Batch Size

Tamás Tardi,Gergo Bogacsovics,Andras Hajdu

Reinforcement learning-based algorithms have been used extensively in recent years due to their flexible nature, good performance, and the increasing number of said algorithms. However, the largest drawback of these techniques remains unsolved, that is, it usually takes a long time for the agents to learn how to solve a given problem. In this work, we outline a novel method that can be used to drastically reduce the training time of current state-of-the-art algorithms like Pr

oximal Policy Optimization (PPO). We evaluate the performance of this approach in a unique environment where we use reinforcement learning to help with a practical astronomical problem: where to place a fixed number of observatory stations in the Solar System to observe space objects (e.g. asteroids) as permanently as possible. That is, the reward in this scenario corresponds to the total coverage of the trajectories of these objects. We apply noisy evaluation for calculating the reward to speed up the training, which technique has already been efficiently applied in stochastic optimization. Namely, we allow the incorporation of some additional noise in the reward function in the form of a hypothesis term and a varying mini-batch size. However, in order to follow the theoretical guidelines, both of them are forced to vanish during training to let the noise converge to zero. Our experimental results show that using this approach we can reduce the training time remarkably, even by 75%.
****

Understanding Train-Validation Split in Meta-Learning with Neural Networks
Xinzhe Zuo,Zixiang Chen,Huaxiu Yao,Yuan Cao,Quanquan Gu

The goal of meta-learning is to learn a good prior model from a collection of tasks such that the learned prior is able to adapt quickly to new tasks without accessing many data from the new tasks. A common practice in meta-learning is to perform a train-validation split on each task, where the training set is used for adapting the model parameter to that specific task and the validation set is used for learning a prior model that is shared across all tasks. Despite its success and popularity in multitask learning and few-shot learning, the understanding of the train-validation split is still limited, especially when the neural network models are used. In this paper, we study the benefit of train-validation split for classification problems with neural network models trained by gradient descent. We prove that the train-validation split is necessary to learn a good prior model when the noise in the training sample is large, while the train-train method fails. We validate our theory by conducting experiment on both synthetic and real datasets. To the best of our knowledge, this is the first work towards the theoretical understanding of train-validation split in meta-learning with neural networks.
****

Revisiting Robustness in Graph Machine Learning
Lukas Gosch,Daniel Sturm,Simon Geisler,Stephan Günnemann

Many works show that node-level predictions of Graph Neural Networks (GNNs) are unrobust to small, often termed adversarial, changes to the graph structure. However, because manual inspection of a graph is difficult, it is unclear if the studied perturbations always preserve a core assumption of adversarial examples: that of unchanged semantic content. To address this problem, we introduce a more principled notion of an adversarial graph, which is aware of semantic content change. Using Contextual Stochastic Block Models (CSBMs) and real-world graphs, our results suggest: $i)$ for a majority of nodes the prevalent perturbation models include a large fraction of perturbed graphs violating the unchanged semantics assumption; $ii)$ surprisingly, all assessed GNNs show over-robustness - that is robustness beyond the point of semantic change. We find this to be a complementary phenomenon to adversarial examples and show that including the label-structure of the training graph into the inference process of GNNs significantly reduces over-robustness, while having a positive effect on test accuracy and adversarial robustness. Theoretically, leveraging our new semantics-aware notion of robustness, we prove that there is no robustness-accuracy tradeoff for inductively classifying a newly added node.
****

Variational Information Pursuit for Interpretable Predictions
Aditya Chattopadhyay,Kwan Ho Ryan Chan,Benjamin David Haeffele,Donald Geman,Rene Vidal

There is a growing interest in the machine learning community in developing predictive algorithms that are interpretable by design. To this end, recent work proposes to sequentially ask interpretable queries about data until a high confidence prediction can be made based on the answers obtained (the history). To promot

e short query-answer chains, a greedy procedure called Information Pursuit (IP) is used, which adaptively chooses queries in order of information gain. Generative models are employed to learn the distribution of query-answers and labels, which is in turn used to estimate the most informative query. However, learning and inference with a full generative model of the data is often intractable for complex tasks. In this work, we propose Variational Information Pursuit (V-IP), a variational characterization of IP which bypasses the need to learn generative models. V-IP is based on finding a query selection strategy and a classifier that minimize the expected cross-entropy between true and predicted labels. We prove that the IP strategy is the optimal solution to this problem. Therefore, instead of learning generative models, we can use our optimal strategy to directly pick the most informative query given any history. We then develop a practical algorithm by defining a finite-dimensional parameterization of our strategy and classifier using deep networks and train them end-to-end using our objective. Empirically, V-IP is 10-100x faster than IP on different Vision and NLP tasks with competitive performance. Moreover, V-IP finds much shorter query chains when compared to reinforcement learning which is typically used in sequential-decision-making problems. Finally, we demonstrate the utility of V-IP on challenging tasks like medical diagnosis where the performance is far superior to the generative modeling approach.
****

Grammar-Induced Geometry for Data-Efficient Molecular Property Prediction
Minghao Guo,Veronika Thost,Samuel Song,Adithya Balachandran,Payel Das,Jie Chen,Wojciech Matusik
The prediction of molecular properties is a crucial task in the field of material and drug discovery. The potential benefits of using deep learning techniques are reflected in the wealth of recent literature. Still, these techniques are faced with a common challenge in practice: Labeled data are limited by the cost of manual extraction from literature and laborious experimentation. In this work, we propose a data-efficient property predictor by utilizing a learnable hierarchical molecular grammar that can generate molecules from grammar production rules. Such a grammar induces an explicit geometry of the space of molecular graphs, which provides an informative prior on molecular structural similarity. The property prediction is performed using graph neural diffusion over the grammar-induced geometry. On both small and large datasets, our evaluation shows that this approach outperforms a wide spectrum of baselines, including supervised and pre-trained graph neural networks. We include a detailed ablation study and further analysis of our solution, showing its effectiveness in cases with extremely limited data (only ${\sim}100$ samples), and its extension to application in molecular generation.
****

EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression
Kaja Gruntkowska,Alexander Tyurin,Peter Richtárik
The starting point of this paper is the discovery of a novel and simple error-feedback mechanism, which we call EF21-P, for dealing with the error introduced by a contractive compressor. Unlike all prior works on error feedback, where compression and correction operate in the dual space of gradients, our mechanism operates in the primal space of models. While we believe that EF21-P may be of interest in many situations where it is often advantageous to perform model perturbation prior to the computation of the gradient (e.g., randomized smoothing and generalization), in this work we focus our attention on its use as a key building block in the design of communication-efficient distributed optimization methods supporting bidirectional compression. In particular, we employ EF21-P as the mechanism for compressing and subsequently error-correcting the model broadcast by the server to the workers. By combining EF21-P with suitable methods performing worker-to-server compression, we obtain novel methods supporting bidirectional compression and enjoying new state-of-the-art theoretical communication complexity for convex and nonconvex problems. For example, our bounds are the first that manage to decouple the variance/error coming from the workers-to-server and s

erver-to-workers compression, transforming a multiplicative dependence to an additive one. In the convex regime, we obtain the first bounds that match the theoretical communication complexity of gradient descent. Even in this convex regime, our algorithms work with biased gradient estimators, which is non-standard and requires new proof techniques that may be of independent interest. Finally, our theoretical results are corroborated through suitable experiments.

****

Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints

Aran Komatsuzaki,Joan Puigcerver,James Lee-Thorp,Carlos Riquelme Ruiz,Basil Mustafa,Joshua Ainslie,Yi Tay,Mostafa Dehghani,Neil Houlsby

Training large, deep neural networks to convergence can be prohibitively expensive. As a result, often only a small selection of popular, dense models are reused across different contexts and tasks. Increasingly, sparsely activated models, which seek to decouple model size from computation costs, are becoming an attractive alternative to dense models. Although more efficient in terms of quality and computation cost, sparse models remain data-hungry and costly to train from scratch in the large scale regime. In this work, we propose sparse upcycling -- a simple way to reuse sunk training costs by initializing a sparsely activated Mixture-of-Experts model from a dense checkpoint. We show that sparsely upcycled T5 Base, Large, and XL language models and Vision Transformer Base and Large models, respectively, significantly outperform their dense counterparts on SuperGLUE and ImageNet, using only ~50% of the initial dense pretraining sunk cost. The upcycled models also outperform sparse models trained from scratch on 100% of the initial dense pretraining computation budget.

****

Lossless Adaptation of Pretrained Vision Models For Robotic Manipulation

Mohit Sharma,Claudio Fantacci,Yuxiang Zhou,Skanda Koppula,Nicolas Heess,Jon Scholz,Yusuf Aytar

Recent works have shown that large models pretrained on common visual learning tasks can provide useful representations for a wide range of specialized perception problems, as well as a variety of robotic manipulation tasks. While prior work on robotic manipulation has predominantly used frozen pretrained features, we demonstrate that in robotics this approach can fail to reach optimal performance, and that fine-tuning of the full model can lead to significantly better results. Unfortunately, fine-tuning disrupts the pretrained visual representation, and causes representational drift towards the fine-tuned task thus leading to a loss of the versatility of the original model. We introduce a method for lossless adaptation to address this shortcoming of classical fine-tuning. We demonstrate that appropriate placement of our parameter efficient adapters can significantly reduce the performance gap between frozen pretrained representations and full end-to-end fine-tuning without changes to the original representation and thus preserving original capabilities of the pretrained model. We perform a comprehensive investigation across three major model architectures (ViTs, NFNets, and ResNets), supervised (ImageNet-1K classification) and self-supervised pretrained weights (CLIP, BYOL, Visual MAE) in three manipulation task domains and 35 individual tasks, and demonstrate that our claims are strongly validated in various settings. Please see real world videos at https://sites.google.com/view/robo-adapters

****

Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

Margaret Li,Suchin Gururangan,Tim Dettmers,Mike Lewis,Tim Althoff,Noah A. Smith,Luke Zettlemoyer

We present Branch-Train-Merge (BTM), a communication-efficient algorithm for embarrassingly parallel training of large language models (LLMs). We show it is possible to independently train subparts of a new class of LLMs on different subsets of the data, eliminating the massive multi-node synchronization currently required to train LLMs. BTM learns a set of independent Expert LMs (ELMs), each specialized to a different textual domain, such as scientific or legal text. These ELMs can be added and removed to update data coverage, ensembled to generalize to new domains, or averaged to collapse back to a single LM for efficient inference. New ELMs are learned by branching from (mixtures of) ELMs in the current se t, further training on new domains, and then merging the resulting models back into the set for future use. Experiments show that BTM improves in- and out-of-domain perplexities as compared to GPT-style Transformer LMs, when controlling for training cost. Through extensive analysis, we show that these results are robust to different ELM initialization schemes, but require expert domain specialization; ensembles with random data splits do not perform well. Our results suggest that aggressive parallelism could be used to efficiently scale larger LMs in future work.

****

Logical Message Passing Networks with One-hop Inference on Atomic Formulas

Zihao Wang,Yangqiu Song,Ginny Wong,Simon See

Complex Query Answering (CQA) over Knowledge Graphs (KGs) has attracted a lot of attention to potentially support many applications. Given that KGs are usually incomplete, neural models are proposed to answer the logical queries by parameterizing set operators with complex neural networks. However, such methods usually train neural set operators with a large number of entity and relation embeddings from the zero, where whether and how the embeddings or the neural set operators contribute to the performance remains not clear. In this paper, we propose a simple framework for complex query answering that decomposes the KG embeddings from neural set operators. We propose to represent the complex queries into the query graph. On top of the query graph, we propose the Logical Message Passing Neural Network (LMPNN) that connects the local one-hop inferences on atomic formulas to the global logical reasoning for complex query answering. We leverage existing effective KG embeddings to conduct one-hop inferences on atomic formulas, the results of which are regarded as the messages passed in LMPNN. The reasoning process over the overall logical formulas is turned into the forward pass of LMPNN that incrementally aggregates local information to finally predict the answers' embeddings. The complex logical inference across different types of queries will then be learned from training examples based on the LMPNN architecture. Theoretically, our query-graph represenation is more general than the prevailing operator-tree formulation, so our approach applies to a broader range of complex KG queries. Empirically, our approach yields the new state-of-the-art neural CQA model. Our research bridges the gap between complex KG query answering tasks and the long-standing achievements of knowledge graph representation learning. Our implementation can be found at https://github.com/HKUST-KnowComp/LMPNN.

****

Noise-Robust De-Duplication at Scale

Emily Silcock,Luca D'Amico-Wong,Jinglin Yang,Melissa Dell

Identifying near duplicates within large, noisy text corpora has a myriad of applications that range from de-duplicating training datasets, reducing privacy risk, and evaluating test set leakage, to identifying reproduced news articles and literature within large corpora. Across these diverse applications, the overwhelming majority of work relies on N-grams. Limited efforts have been made to evaluate how well N-gram methods perform, in part because it is unclear how one could create an unbiased evaluation dataset for a massive corpus. This study uses the unique timeliness of historical news wires to create a 27,210 document dataset, with 122,876 positive duplicate pairs, for studying noise-robust de-duplication. The time-sensitivity of news makes comprehensive hand labelling feasible - despite the massive overall size of the corpus - as duplicates occur within a narrow date range. The study then develops and evaluates a range of de-duplication methods: hashing and N-gram overlap (which predominate in the literature), a contrastively trained bi-encoder, and a ``re-rank'' style approach combining a bi- and cross-encoder. The neural approaches significantly outperform hashing and N-gram overlap. We show that the bi-encoder scales well, de-duplicating a 10 million article corpus on a single GPU card in a matter of hours. We also apply our pre-trained model to the RealNews and patent portions of C4 (Colossal Clean Crawled Corpus), illustrating that a neural approach can identify many near duplicates missed by hashing, in the presence of various types of noise. The public release of our NEWS-COPY de-duplication dataset, codebase, and the pre-trained models will facilitate further research and applications.

****

P2PRISM - Peer to peer learning with individual prism for secure aggregation
Atul Sharma,Wei Chen,Joshua Christian Zhao,Qiang Qiu,Saurabh Bagchi,Somali Chaterji
Federated learning (FL) has made collaboration between nodes possible without explicit sharing of local data. However, it requires the participating nodes to trust the server and its model updates, the server itself being a critical node susceptible to failure and compromise. A loss of trust in the server and a demand to aggregate the model independently for oneself has led decentralized peer-to-peer learning (P2PL) to gain traction lately. In this paper, we highlight the never before exposed vulnerabilities of P2PL towards malicious attacks and how P2PL behaves differently from FL in such a malicious environment. We then present a robust defense - P2PRISM as a secure aggregation protocol for P2PL.

****

Multi-scale Attention for Diabetic Retinopathy Detection in Retinal Fundus Images
Temitope Ibrahim Amosa,Patrick Sebastian,Lila Iznita Izhar,Fatimat Adeola Adekola,Mardiyyah Adeola Salahudeen
The diagnosis and/or grading of diabetic retinopathy (DR) in the retina fundus has traditionally been done by physicians using manual procedures. However, there has been a significant demand for automated eye diagnostic and grading systems due to the constant rise in the number of persons with diabetes over the past few decades. An excellent diagnostic and predictive value for treatment planning exists with automatic DR grading based on retinal fundus pictures. With the majority of the current automated DR grading systems, it is exceedingly challenging to capture significant features because of the minor changes between severity levels. This paper presents a deep learning-based method for automatically assessing diabetic retinopathy in retina fundus pictures. This paper presents a deep learning-based method for automatically assessing diabetic retinopathy in retina fundus pictures. In order to increase the discriminative ability of the retrieved features, we implement a multi-scale attention mechanism within a deep convolutional neural network architecture in this research. Additionally, we provide a brand-new loss function termed modified grading loss that enhances the training convergence of the suggested strategy by taking into account the distance between various grades of distinct DR categories. The suggested technique is trained, validated, and tested using a dataset about diabetic retinopathy that is openly available. The experimental findings are presented to illustrate how well the suggested strategy competes.

****

Blessing from Experts: Super Reinforcement Learning in Confounded Environments
Jiayi Wang,Zhengling Qi,Chengchun Shi
We introduce super reinforcement learning in the batch setting, which takes the observed action as input for enhanced policy learning. In the presence of unmeasured confounders, the recommendations from human experts recorded in the observed data allow us to recover certain unobserved information. Including this information in the policy search, the proposed super reinforcement learning will yield a super policy that is guaranteed to outperform both the standard optimal policy and the behavior one (e.g., the expert’s recommendation). Furthermore, to address the issue of unmeasured confounding in finding super-policies, a number of non-parametric identification results are established. Finally, we develop two super-policy learning algorithms and derive their corresponding finite-sample regret guarantees.

****

Reinforcement Learning for Bandits with Continuous Actions and Large Context Spaces
Paul Duckworth,Bruno Lacerda,Katherine Vallis,Nick Hawes
We consider the challenging scenario of contextual bandits with continuous actions and large input ``context'' spaces, e.g. images. We posit that by modifying reinforcement learning (RL) algorithms for continuous control, we can outperform hand-crafted contextual bandit algorithms for continuous actions on standard ben

chmark datasets, i.e. vector contexts. We demonstrate that parametric policy networks outperform recently published tree-based policies in both average regret and costs on held-out samples. Furthermore, in contrast to previous work, we successfully demonstrate that RL algorithms can generalise contextual bandit problems with continuous actions to large context spaces. We obtain state-of-the-art performance using RL and significantly outperform previous methods on image contexts. Lastly, we introduce a new contextual bandits domain with multi-dimensional continuous action space and image context.

****

Explanation Uncertainty with Decision Boundary Awareness

Davin Hill,Aria Masoomi,Sandesh Ghimire,Max Torop,Jennifer Dy

Post-hoc explanation methods have become increasingly depended upon for understanding black-box classifiers in high-stakes applications, precipitating a need for reliable explanations. While numerous explanation methods have been proposed, recent works have shown that many existing methods can be inconsistent or unstable. In addition, high-performing classifiers are often highly nonlinear and can exhibit complex behavior around the decision boundary, leading to brittle or misleading local explanations. Therefore, there is an impending need to quantify the uncertainty of such explanation methods in order to understand when explanations are trustworthy. We introduce a novel uncertainty quantification method parameterized by a Gaussian Process model, which combines the uncertainty approximation of existing methods with a novel geodesic-based similarity which captures the complexity of the target black-box decision boundary. The proposed framework is highly flexible—it can be used with any black-box classifier and feature attribution method to amortize uncertainty estimates for explanations. We show theoretically that our proposed geodesic-based kernel similarity increases with the complexity of the decision boundary. Empirical results on multiple tabular and image datasets show that our decision boundary-aware uncertainty estimate improves understanding of explanations as compared to existing methods

****

SARNET: SARCASM VS TRUE-HATE DETECTION NETWORK

Harsh Mittal,Kartikeya Singh Chauhan,Anil Singh Parihar,Kavinder Singh,Ashutosh Pandey

At times hate speech detection classifiers miss the context of a sentence and flag a sarcastic tweet incorrectly. To tackle this problem by emphasising on the context of a tweet we propose SarNet. SarNet is a two-fold deep learning based model which follows a quasi-ternary labelling strategy and contextually classifies a tweet as hate, sarcastic or neither. The first module of SarNet is an ANN-BiLSTM based Pyramid Network used to calculate the hate and sarcastic probabilities of a sentence. The second module of the SarNet is the Nash Equalizer which stems from the concept of game theory and prisoner’s dilemma. It treats hate and sarcasm as two prisoners. A payoff matrix is constructed to calculate the true hate of the tweet. True hate considers the hate part of a tweet excluding the sarcastic part of the tweet. Thus, this gives a true estimate of the hate content in a tweet thereby decreasing the number of sarcastic tweets being falsely flagged as hate. Our proposed model is trained on state-of-the-art hate speech and sarcasm datasets in the English language. The precision, recall and F1 score of our proposed model is 0.93, 0.84 and 0.88 respectively. Comparison with state-of-the-art architectures demonstrated better performance of SarNet by a significant margin.

****

Learning Portable Skills by Identifying Generalizing Features with an Attention-Based Ensemble

Anita Taosheng De Mello Koch,Zhiyuan Zhou,Akhil Bagaria,Haotian Fu,Cameron Allen,George Konidaris

The ability to rapidly generalize is crucial for reinforcement learning to be practical in real-world tasks. However, generalization is complicated by the fact that, in many settings, some state features reliably support generalization while others do not. We consider the problem of learning generalizable policies and skills (in the form of options) by identifying feature sets that generalize acro ss instances. We propose an attention-ensemble approach, where a collection of minimally overlapping feature masks is learned, each of which individually maximizes performance on the source instance. Subsequent tasks are instantiated using the ensemble, and transfer performance is used to update the estimated probability that each feature set will generalize in the future. We show that our approach leads to fast policy generalization for eight tasks in the Procgen benchmark. We then show its use in learning portable options in Montezuma's Revenge, where it is able to generalize skills learned in the first screen to the remainder of the game.

****

Few-shot Backdoor Attacks via Neural Tangent Kernels
Jonathan Hayase,Sewoong Oh

In a backdoor attack, an attacker injects corrupted examples into the training set. The goal of the attacker is to cause the final trained model to predict the attacker's desired target label when a predefined trigger is added to test inputs. Central to these attacks is the trade-off between the success rate of the attack and the number of corrupted training examples injected. We pose this attack as a novel bilevel optimization problem: construct strong poison examples that maximize the attack success rate of the trained model. We use neural tangent kernels to approximate the training dynamics of the model being attacked and automatically learn strong poison examples. We experiment on subclasses of CIFAR-10 and ImageNet with WideResNet-34 and ConvNeXt architectures on periodic and patch trigger attacks and show that NTBA-designed poisoned examples achieve, for example, an attack success rate of 90% with ten times smaller number of poison examples injected compared to the baseline. We provided an interpretation of the NTBA-designed attacks using the analysis of kernel linear regression. We further demonstrate a vulnerability in overparametrized deep neural networks, which is revealed by the shape of the neural tangent kernel.

****

Quantitative Universal Approximation Bounds for Deep Belief Networks
Julian Sieber,Johann Gehringer

We show that deep belief networks with binary hidden units can approximate any multivariate probability density under very mild integrability requirements on the parental density of the visible nodes. The approximation is measured in the L^q-norm for $q\in[1,\infty]$ ($q=\infty$ corresponding to the supremum norm) and in Kullback-Leibler divergence. Furthermore, we establish sharp quantitative bounds on the approximation error in terms of the number of hidden units.

****

Hyperparameter Optimization through Neural Network Partitioning
Bruno Kacper Mlodozeniec,Matthias Reisser,Christos Louizos

Well-tuned hyperparameters are crucial for obtaining good generalization behavior in neural networks. They can enforce appropriate inductive biases, regularize the model and improve performance --- especially in the presence of limited data. In this work, we propose a simple and efficient way for optimizing hyperparameters inspired by the marginal likelihood, an optimization objective that requires no validation data. Our method partitions the training data and a neural network model into K data shards and parameter partitions, respectively. Each partition is associated with and optimized only on specific data shards. Combining these partitions into subnetworks allows us to define the "out-of-training-sample" loss of a subnetwork, i.e., the loss on data shards unseen by the subnetwork, as the objective for hyperparameter optimization. We demonstrate that we can apply this objective to optimize a variety of different hyperparameters in a single training run while being significantly computationally cheaper than alternative methods aiming to optimize the marginal likelihood for neural networks. Lastly, we also focus on optimizing hyperparameters in federated learning, where retraining and cross-validation are particularly challenging.

****

DiscoBAX - Discovery of optimal intervention sets in genomic experiment design
Clare Lyle,Arash Mehrjou,Pascal Notin,Andrew Jesson,Stefan Bauer,Yarin Gal,Patri

ck Schwab

The discovery of novel therapeutics to cure genetic pathologies relies on the identification of the different genes involved in the underlying disease mechanism. With billions of potential hypotheses to test, an exhaustive exploration of the entire space of potential interventions is impossible in practice. Sample-efficient methods based on active learning or bayesian optimization bear the promise of identifying interesting targets using the least experiments possible. However, genomic perturbation experiments typically rely on proxy outcomes measured in biological model systems that may not completely correlate with the outcome of interventions in humans. In practical experiment design, one aims to find a set of interventions which maximally move a target phenotype via a diverse set of mechanisms in order to reduce the risk of failure in future stages of trials. To that end, we introduce DiscoBAX — a sample-efficient algorithm for the discovery of genetic interventions that maximize the movement of a phenotype in a direction of interest while covering a diverse set of underlying mechanisms. We provide theoretical guarantees on the optimality of the approach under standard assumptions, conduct extensive experiments in synthetic and real-world settings relevant to genomic discovery and demonstrate that DiscoBAX outperforms state-of-the-art active learning and Bayesian optimization methods in this task. Better methods for selecting effective and diverse perturbations in biological systems could enable researchers to potentially discover novel therapeutics for a range of genetically-driven diseases.

****

How to Enable Uncertainty Estimation in Proximal Policy Optimization

Eugene Bykovets,Yannick Metz,Mennatallah El-Assady,Daniel A. Keim,Joachim M. Buhmann

While deep reinforcement learning (RL) agents have showcased strong results across many domains, a major concern is their inherent opaqueness and the safety of such systems in real-world use cases. To overcome these issues, we need agents that can quantify their uncertainty and detect out-of-distribution (OOD) states. Existing uncertainty estimation techniques, like Monte-Carlo Dropout or Deep Ensembles, have not seen widespread adoption in on-policy deep RL. We posit that this is due to two reasons: concepts like uncertainty and OOD states are not well defined compared to supervised learning, especially for on-policy RL methods. Secondly, available implementations and comparative studies for uncertainty estimation methods in RL have been limited. To overcome the first gap, we propose definitions of uncertainty and OOD for Actor-Critic RL algorithms, namely, proximal policy optimization (PPO), and present possible applicable measures. In particular, we discuss the concepts of value and policy uncertainty. The second point is addressed by implementing different uncertainty estimation methods and comparing them across a number of environments. The OOD detection performance is evaluated via a custom evaluation benchmark of in-distribution (ID) and OOD states for various RL environments. We identify a trade-off between reward and OOD detection performance. To overcome this, we formulate a Pareto optimization problem in which we simultaneously optimize for reward and OOD detection performance. We show experimentally that the recently proposed method of Masksembles strikes a favourable balance among the survey methods, enabling high-quality uncertainty estimation and OOD detection while matching the performance of original RL agents.

****

Joint-Predictive Representations for Multi-Agent Reinforcement Learning

Mingxiao Feng,Wengang Zhou,Yaodong Yang,Houqiang Li

The recent advances in reinforcement learning have demonstrated the effectiveness of vision-based self-supervised learning (SSL). However, the main efforts on this direction have been paid on single-agent setting, making multi-agent reinforcement learning~(MARL) lags thus far. There are two significant obstacles that prevent applying off-the-shelf SSL approaches with MARL on a partially observable multi-agent system : (a) each agent only gets a partial observation, and (b) previous SSL approaches only take consistent temporal representations into account, while ignoring the characterization that captures the interaction and fusion among agents. In this paper, we propose \textbf{M}ulti-\textbf{A}gent \textbf{Jo

}int-Predictive \textbf{R}epresentations~(MAJOR), a novel framework to explore self-supervised learning on cooperative MARL. Specifically, we treat the latent representations of local observations of all agents as the sequence of masked contexts of the global state, and we then learn effective representations by predicting the future latent representations for each agent with the help of the agent-level information interactions in a joint transition model. We have conducted extensive experiments on wide-range MARL environments, including both vision-based and state-based scenarios, and show that our proposed MAJOR achieves superior asymptotic performance and sample efficiency against other state-of-the-art methods.

****

Symmetries, Flat Minima and the Conserved Quantities of Gradient Flow

Bo Zhao,Iordan Ganev,Robin Walters,Rose Yu,Nima Dehmamy

Empirical studies of the loss landscape of deep networks have revealed that many local minima are connected through low-loss valleys. Yet, little is known about the theoretical origin of such valleys. We present a general framework for finding continuous symmetries in the parameter space, which carve out low-loss valleys. Our framework uses equivariances of the activation functions and can be applied to different layer architectures. To generalize this framework to nonlinear neural networks, we introduce a novel set of nonlinear, data-dependent symmetries. These symmetries can transform a trained model such that it performs similarly on new samples, which allows ensemble building that improves robustness under certain adversarial attacks. We then show that conserved quantities associated with linear symmetries can be used to define coordinates along low-loss valleys. The conserved quantities help reveal that using common initialization methods, gradient flow only explores a small part of the global minimum. By relating conserved quantities to convergence rate and sharpness of the minimum, we provide insights on how initialization impacts convergence and generalizability.

****

DP-SGD-LF: Improving Utility under Differentially Private Learning via Layer Freezing

Qiaoyue Tang,Mathias Lécuyer

Differentially Private SGD (DP-SGD) is a widely known substitute for SGD to train deep learning models with privacy guarantees. However, privacy guarantees come at cost in model utility. The key DP-SGD steps responsible for this utility cost are per-sample gradient clipping, which introduces bias, and adding noise to the aggregated (clipped) gradients, which increases the variance of model updates. Inspired by the observation that different layers in a neural network often converge at different rates following a bottom-up pattern, we incorporate layer freezing into DP-SGD to increase model utility at fixed privacy budget. Through theoretical analysis and empirical evidence we show that layer freezing improves model utility, by reducing both the bias and variance introduced by gradient clipping and noising. These improvements in turn lead to better model accuracy, and empirically generalize over multiple datasets, models, and privacy budgets.

****

Explainability as statistical inference

Hugo Henri Joseph Senetaire,Damien Garreau,Jes Frellsen,Pierre-Alexandre Mattei

A wide variety of model explanation approaches have been proposed in recent years, all guided by very different rationales and heuristics. In this paper, we take a new route and cast interpretability as a statistical inference problem. We propose a general deep probabilistic model designed to produce interpretable predictions. The model’s parameters can be learned via maximum likelihood, and the method can be adapted to any predictor network architecture, and any type of prediction problem. Our method is a case of amortized interpretability models, where a neural network is used as a selector to allow for fast interpretation at inference time. Several popular interpretability methods are shown to be particular cases of regularised maximum likelihood for our general model. We propose new datasets with ground truth selection which allow for the evaluation of the features importance map. Using these datasets, we show experimentally that using multip

le imputation provides more reasonable interpretation.
****

Concept-based Explanations for Out-of-Distribution Detectors
Jihye Choi,Jayaram Raghuram,Ryan Feng,Jiefeng Chen,Somesh Jha,Atul Prakash
Out-of-distribution (OOD) detection plays a crucial role in ensuring the safe deployment of deep neural network (DNN) classifiers.
While a myriad of methods have focused on improving the performance of OOD detectors, a critical gap remains in interpreting their decisions.
We help bridge this gap by providing explanations for OOD detectors based on learned high-level concepts.
We first propose two new metrics for assessing the effectiveness of a particular set of concepts for explaining OOD detectors: 1) $\textit{detection completeness}$, which quantifies the sufficiency of concepts for explaining an OOD-detector's decisions, and 2) $\textit{concept separability}$, which captures the distributional separation between in-distribution and OOD data in the concept space.
Based on these metrics, we propose a framework for learning a set of concepts that satisfy the desired properties of detection completeness and concept separability, and demonstrate the framework's effectiveness in providing concept-based explanations for diverse OOD detection techniques.
We also show how to identify prominent concepts that contribute to the detection results via a modified Shapley value-based importance score.
****

FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels
Guillaume Staerman,Cédric Allain,Alexandre Gramfort,Thomas Moreau
Temporal point processes (TPP) are a natural tool for modeling event-based data. Among all TPP models, Hawkes processes have proven to be the most widely used, mainly due to their simplicity and computational ease when considering exponential or non-parametric kernels. Although non-parametric kernels are an option, such models require large datasets. While exponential kernels are more data efficient and relevant for certain applications where events immediately trigger more events, they are ill-suited for applications where latencies need to be estimated, such as in neuroscience. This work aims to offer an efficient solution to TPP inference using general parametric kernels with finite support. The developed solution consists of a fast L2 gradient-based solver leveraging a discretized version of the events. After supporting the use of discretization theoretically, the statistical and computational efficiency of the novel approach is demonstrated through various numerical experiments. Finally, the effectiveness of the method is evaluated by modeling the occurrence of stimuli-induced patterns from brain signals recorded with magnetoencephalography (MEG). Given the use of general parametric kernels, results show that the proposed approach leads to a more plausible estimation of pattern latency compared to the state-of-the-art.

****

Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees
Swarnadeep Saha,Shiyue Zhang,Peter Hase,Mohit Bansal
Current abstractive summarization models either suffer from a lack of clear interpretability or provide incomplete rationales by only highlighting parts of the source document. To this end, we propose the Summarization Program (SP), an interpretable modular framework consisting of an (ordered) list of binary trees, each encoding the step-by-step generative process of an abstractive summary sentence from the source document. A Summarization Program contains one root node per summary sentence, and a distinct tree connects each summary sentence (root node) to the document sentences (leaf nodes) from which it is derived, with the connecting nodes containing intermediate generated sentences. Edges represent different modular operations involved in summarization such as sentence fusion, compression, and paraphrasing. We first propose an efficient best-first search method over neural modules, SP-Search that identifies SPs for human summaries by directly optimizing for ROUGE scores. Next, using these programs as automatic supervisio

n, we propose seq2seq models that generate Summarization Programs, which are then executed to obtain final summaries. We demonstrate that SP-Search effectively represents the generative process behind human summaries using modules that are typically faithful to their intended behavior. We also conduct a simulation study to show that Summarization Programs improve the interpretability of summarization models by allowing humans to better simulate model reasoning. Summarization Programs constitute a promising step toward interpretable and modular abstractive summarization, a complex task previously addressed primarily through blackbox end-to-end neural systems.

****

Planning with Large Language Models for Code Generation

Shun Zhang,Zhenfang Chen,Yikang Shen,Mingyu Ding,Joshua B. Tenenbaum,Chuang Gan

Existing large language model-based code generation pipelines typically use beam search or sampling algorithms during the decoding process. Although the programs they generate achieve high token-matching-based scores, they often fail to compile or generate incorrect outputs. The main reason is that conventional Transformer decoding algorithms may not be the best choice for code generation. In this work, we propose a novel Transformer decoding algorithm, Planning-Guided Transformer Decoding (PG-TD), that uses a planning algorithm to do lookahead search and guide the Transformer to generate better programs. Specifically, instead of simply optimizing the likelihood of the generated sequences, the Transformer makes use of a planner that generates candidate programs and tests them on public test cases. The Transformer can therefore make more informed decisions and generate tokens that will eventually lead to higher-quality programs. We also design a mechanism that shares information between the Transformer and the planner to make our algorithm computationally efficient. We empirically evaluate our framework with several large language models as backbones on public coding challenge benchmarks, showing that 1) it can generate programs that consistently achieve higher performance compared with competing baseline methods; 2) it enables controllable code generation, such as concise codes and highly-commented codes by optimizing modified objective.

****

Unleash Model Capacity for Universal Dense Retrieval by Task Specialty Optimization

Wenzheng Zhang,Chenyan Xiong,Karl Stratos,Arnold Overwijk

Universal dense retrieval, with one unified representation space to empower various retrieval scenarios, has many appealing advantages in simplicity, efficiency, and potential to break echo chambers with cross-scenario information access. However, standard multi-task trained dense retrievers often fail to meet the accuracy of scenario-specific models. In this paper, we analyze the multi-task learning in universal retrieval and show that the model capacity is not the main bottleneck. It is the optimization failed to fully utilize the network parameters to capture task-specific signals. This motivated our development of TACO-DR, which conducts multi-task learning for universal retrieval with TAsk speCialty Optimization. TACO-DR dynamically adjusts the learning rate for each parameter regrading each task based on its task-specific sensitivity, to encourage parameters to better capture task specific signals. On the KILT benchmark, TACO-DR outperforms various multi-task learning methods and achieves better overall accuracy than single-task models. Our analysis shows that TACO-DR better utilizes the model capacity with more task-specific parameters. Our code and model checkpoints will be open-sourced.

****

Training Equilibria in Reinforcement Learning

Lauro Langosco,David Krueger,Adam Gleave

In partially observable environments, reinforcement learning algorithms such as policy gradient and Q-learning may have multiple equilibria---policies that are stable under further training---and can converge to policies that are strictly suboptimal.

Prior work blames insufficient exploration, but suboptimal equilibria can arise despite full exploration and other favorable circumstances like a flexible polic

y parametrization.
We show theoretically that the core problem is that in partially observed environments, an agent's past actions induce a distribution on hidden states.
Equipping the policy with memory helps it model the hidden state and leads to convergence to a higher reward equilibrium, \emph{even when there exists a memoryless optimal policy}.
Experiments show that
policies with insufficient memory tend to learn to use the environment as auxiliary memory,and parameter noise helps policies escape suboptimal equilibria.
****

Hebbian Deep Learning Without Feedback
Adrien Journé,Hector Garcia Rodriguez,Qinghai Guo,Timoleon Moraitis
Recent approximations to backpropagation (BP) have mitigated many of BP's computational inefficiencies and incompatibilities with biology, but important limitations still remain. Moreover, the approximations significantly decrease accuracy in benchmarks, suggesting that an entirely different approach may be more fruitful. Here, grounded on recent theory for Hebbian learning in soft winner-take-all networks, we present multilayer SoftHebb, i.e. an algorithm that trains deep neural networks, without any feedback, target, or error signals. As a result, it achieves efficiency by avoiding weight transport, non-local plasticity, time-locking of layer updates, iterative equilibria, and (self-) supervisory or other feedback signals – which were necessary in other approaches. Its increased efficiency and biological compatibility do not trade off accuracy compared to state-of-the-art bio-plausible learning, but rather improve it. With up to five hidden layers and an added linear classifier, accuracies on MNIST, CIFAR-10, STL-10, and ImageNet, respectively reach 99.4%, 80.3%, 76.2%, and 27.3%. In conclusion, SoftHebb shows with a radically different approach from BP that Deep Learning over few layers may be plausible in the brain and increases the accuracy of bio-plausible machine learning. Code is available at https://github.com/NeuromorphicComputing/SoftHebb.
****

A Simulation-based Framework for Robust Federated Learning to Training-time Attacks
Wanyun Xie,Thomas Pethick,Ali Ramezani-Kebrya,Volkan Cevher
Well-known robust aggregation schemes in federated learning (FL) are shown to be vulnerable to an informed adversary who can tailor training-time attacks [Fang et al., Xie et al.]. We frame robust distributed learning problem as a game between a server and an adversary that is able to optimize strong training-time attacks. We introduce RobustTailor, a simulation-based framework that prevents the adversary from being omniscient. The simulated game we propose enjoys theoretical guarantees through a regret analysis. RobustTailor improves robustness to training-time attacks significantly while preserving almost the same privacy guarantees as standard robust aggregation schemes in FL. Empirical results under challenging attacks show that RobustTailor performs similar to an upper bound with perfect knowledge of honest clients.
****

Key Design Choices for Double-transfer in Source-free Unsupervised Domain Adaptation
Andrea Maracani,Raffaello Camoriano,Elisa Maiettini,Davide Talon,Lorenzo Rosasco,Lorenzo Natale
Fine-tuning and Domain Adaptation emerged as effective strategies for efficiently transferring deep learning models to new target tasks. However, target domain labels are not accessible in many real-world scenarios. This led to the development of Unsupervised Domain Adaptation (UDA) methods, which only employ unlabeled target samples. Furthermore, efficiency and privacy requirements may also prevent the use of source domain data during the adaptation stage. This particularly challenging setting, known as Source-free Unsupervised Domain Adaptation (SF-UDA), is still understudied. In this paper, we systematically analyze the impact of the main design choices in SF-UDA through a large-scale empirical study on 500 models and 74 domain pairs. We identify the normalization approach, pre-training

strategy, and backbone architecture as the most critical factors. Based on our observations, we propose recipes to best tackle SF-UDA scenarios. Moreover, we show that SF-UDA performs competitively also beyond standard benchmarks and backbone architectures, performing on par with UDA at a fraction of the data and computational cost. Experimental data and code will be released upon acceptance.
****

PALM: Preference-based Adversarial Manipulation against Deep Reinforcement Learning

Fengshuo Bai,Runze Liu,Yaodong Yang,Yali Du

To improve the robustness of DRL agents, it is important to study their vulnerability under adversarial attacks that would lead to extreme behaviors desired by adversaries. Preference-based RL (PbRL) aims for learning desired behaviors with human preferences. In this paper, we propose PALM, a preference-based adversarial manipulation method against DRL agents which adopts human preferences to perform targeted attacks with the assistance of an intention policy and a weighting function. The intention policy is trained based on the PbRL framework to guide the adversarial policy to mitigate restrictions of the victim policy during exploration, and the weighting function learns weight assignment to improve the performance of the adversarial policy. Theoretical analysis demonstrates that PALM converges to critical points under some mild conditions. Empirical results on a few manipulation tasks of Meta-world show that PALM exceeds the performance of state-of-the-art adversarial attack methods under the targeted setting. Additionally, we show the vulnerability of the offline RL agents by fooling them into behaving as human desires on several Mujoco tasks. Our code and videos are available in https://sites.google.com/view/palm-adversarial-attack.
****

Equivariance-aware Architectural Optimization of Neural Networks

Kaitlin Maile,Dennis George Wilson,Patrick Forré

Incorporating equivariance to symmetry groups as a constraint during neural network training can improve performance and generalization for tasks exhibiting those symmetries, but such symmetries are often not perfectly nor explicitly present. This motivates algorithmically optimizing the architectural constraints imposed by equivariance. We propose the equivariance relaxation morphism, which preserves functionality while reparameterizing a group equivariant layer to operate with equivariance constraints on a subgroup, as well as the $[G]$-mixed equivariant layer, which mixes layers constrained to different groups to enable within-layer equivariance optimization. We further present evolutionary and differentiable neural architecture search (NAS) algorithms that utilize these mechanisms respectively for equivariance-aware architectural optimization. Experiments across a variety of datasets show the benefit of dynamically constrained equivariance to find effective architectures with approximate equivariance.
****

Unsupervised Non-Parametric Signal Separation Using Bayesian Neural Networks

Alexander V Belikov,Alessandro Montanari,Emmanuel Simon Moulin

Bayesian neural networks (BNN) take the best from two worlds: the one of flexible and scalable neural networks and the one of probabilistic graphical models, the latter allowing for probabilistic interpretation of inference results.

We make one extra step towards unification of these two domains and render BNN as an elementary unit of abstraction in the framework of probabilistic modeling, which allows us to promote well-known distributions to distribution fields.

We use transformations to obtain field versions of several popular distributions and demonstrate the utility of our approach on the problem of signal/background separation.

Starting from prior knowledge that a certain region of space contains predominantly one of the components, in an unsupervised and non-parametric manner, we recover the representation of both previously unseen components as well as their proportions.
****

SPIDER: Searching Personalized Neural Architecture for Federated Learning

Erum Mushtaq,Chaoyang He,Jie Ding,Salman Avestimehr

Federated learning (FL) is an efficient learning framework that assists distributed machine learning when data cannot be shared with a centralized server. Recent advancements in FL use predefined architecture-based learning for all the clients. However, given that clients' data are invisible to the server and data distributions are non-identical across clients, a predefined architecture discovered in a centralized setting may not be an optimal solution for all the clientsin FL. Motivated by this challenge, we introduce SPIDER, an algorithmic framework that aims to Search Personalized neural architecture for feDERated learning. SPIDER is designed based on two unique features: (1) alternately optimizing one architecture-homogeneous global model (Supernet) in a generic FL manner and one architecture-heterogeneous local model that is connected to the global model by weight-sharing-based regularization (2) achieving architecture-heterogeneous local model by an operation-level perturbation based neural architecture search method. Experimental results demonstrate that SPIDER outperforms other state-of-the-art personalization methods with much fewer times of hyperparameter tuning.
****
On Gradient Descent Convergence beyond the Edge of Stability
Lei Chen,Joan Bruna
Gradient Descent (GD) is a powerful workhorse of modern machine learning thanks to its scalability and efficiency in high-dimensional spaces. Its ability to find local minimisers is only guaranteed for losses with Lipschitz gradients, where it can be seen as a `bona-fide' discretisation of an underlying gradient flow. Yet, many ML setups involving overparametrised models do not fall into this problem class, which has motivated research beyond the so-called ``Edge of Stability'' (EoS), where the step-size crosses the admissibility threshold inversely proportional to the Lipschitz constant above. Perhaps surprisingly, GD has been empirically observed to still converge regardless of local instability and oscillatory behavior.

The incipient theoretical analysis of this phenomena has mainly focused in the overparametrised regime, where the effect of choosing a large learning rate may be associated to a `Sharpness-Minimisation' implicit regularisation within the manifold of minimisers, under appropriate asymptotic limits. In contrast, in this work we directly examine the conditions for such unstable convergence, focusing on simple, yet representative, learning problems. Specifically, we characterize a local condition involving third-order derivatives that stabilizes oscillations of GD above the EoS, and leverage such property in a teacher-student setting, under population loss. Finally, focusing on Matrix Factorization, we establish a non-asymptotic `Local Implicit Bias' of GD above the EoS, whereby quasi-symmetric initializations converge to symmetric solutions --- where sharpness is minimum amongst all minimisers.
****
Synaptic Dynamics Realize First-order Adaptive Learning and Weight Symmetry
Yukun Yang,Peng Li
Gradient-based first-order adaptive optimization methods such as the Adam optimizer are prevalent in training artificial networks, achieving the state-of-the-art results. This work attempts to answer the question whether it is viable for biological neural systems to adopt such optimization methods. To this end, we demonstrate a realization of the Adam optimizer using biologically-plausible mechanisms in synapses. The proposed learning rule has clear biological correspondence, runs continuously in time, and achieves performance to comparable Adam's. In addition, we present a new approach, inspired by the predisposition property of synapses observed in neuroscience, to circumvent the biological implausibility of the weight transport problem in backpropagation (BP). With only local information and no separate training phases, this method establishes and maintains weight symmetry in the forward and backward signaling paths, and is applicable to the proposed biologically plausible Adam learning rule. The aforementioned mechanisms may shed light on the way in which biological synaptic dynamics facilitate learning.
****

FedAvg Converges to Zero Training Loss Linearly: The Power of Overparameterized Multi-Layer Neural Networks
Bingqing Song,Prashant Khanduri,Xinwei Zhang,Jinfeng Yi,Mingyi Hong
Federated Learning (FL) is a distributed learning paradigm that allows multiple clients to learn a joint model by utilizing privately held data at each client. Significant research efforts have been devoted to develop advanced algorithms that deal with the situation where the data at individual clients have different distributions (i.e., the data heterogeneity issue). In this work, we show that data heterogeneity can be dealt from a different perspective. That is, by utilizing a certain overparameterized multi-layer neural network at each client, even the vanilla FedAvg (a.k.a. the Local SGD) algorithm can accurately optimize the training problem. Specifically, when each client has a neural network with one wide layer of size N (where N is the number of total training samples), followed by layers of smaller widths, FedAvg converges linearly to a solution that achieves (almost) zero training loss, without requiring any assumptions on the data distributions at each client. To our knowledge, this is the first work that demonstrates such resilience to data heterogeneity
for FedAvg when trained on multi-layer neural networks. Our experiments also confirm that, neural network of large size can achieve better and more stable performance for FL problems.
****
Robust Graph Representation Learning via Predictive Coding
Billy Byiringiro,Tommaso Salvatori,Thomas Lukasiewicz
Graph neural networks have recently shown outstanding results in diverse types of tasks in machine learning, providing interdisciplinary state-of-the-art performance on structured data. However, they have been proved to be vulnerable to imperceptible adversarial attacks and shown to be unfit for out-of-distribution generalisation.
Here, we address this problem by introducing a novel message-passing scheme based on the theory of predictive coding, an energy-based alternative to back-propagation that has its roots in neuroscience.
As both graph convolution and predictive coding can be seen as low-pass filtering mechanisms, we postulate that predictive coding adds a second efficient filter to the messaging passing process which enhances the robustness of the learned representation. Through an extensive set of experiments, we show that the proposed model attains comparable performance to its graph convolution network counterpart, delivering strictly better performance on inductive tasks. Most importantly, we show that the energy minimization enhances the robustness of the produced presentation and can be leveraged to further calibrate our models and provide representations that are more robust against advanced graph adversarial attacks.

****
Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time
Jun-Kun Wang,Andre Wibisono
Hamiltonian Monte Carlo (HMC) is a popular method in sampling. While there are quite a few works of studying this method on various aspects, an interesting question is how to choose its integration time to achieve acceleration. In this work, we consider accelerating the process of sampling from a distribution $\pi(x) \propto \exp(-f(x))$ via HMC via time-varying integration time. When the potential f is L-smooth and m-strongly convex, i.e. for sampling from a log-smooth and strongly log-concave target distribution π, it is known that under a constant integration time, the number of iterations that ideal HMC takes to get an ϵ Wasserstein-2 distance to the target π is $O(\kappa \log \frac{1}{\epsilon})$, where $\kappa := \frac{L}{m}$ is the condition number. We propose a scheme of time-varying integration time based on the roots of Chebyshev polynomials. We show that in the case of quadratic potential f, i.e. when the target π is a Gaussian distribution, ideal HMC with this choice of integration time only takes $O(\sqrt{\kappa} \log \frac{1}{\epsilon})$ number of iterations to reach Wasserstein-2 distance less than ϵ; this improvement on the dependence on condition number is akin to acceleration in optimization. The desi

gn and analysis of HMC with the proposed integration time is built on the tools of Chebyshev polynomials. Experiments find the advantage of adopting our scheme of time-varying integration time even for sampling from distributions with smooth strongly convex potentials that are not quadratic.

****

Multi-Hypothesis 3D human pose estimation metrics favor miscalibrated distributions

Pawe■ A. Pierzchlewicz,R. James Cotton,Mohammad Bashiri,Fabian H. Sinz

Due to depth ambiguities and occlusions, lifting 2D poses to 3D is a highly ill-posed problem. Well-calibrated distributions of possible poses can make these ambiguities explicit and preserve the resulting uncertainty for downstream tasks. This study shows that previous attempts, which account for these ambiguities via multiple hypotheses generation, produce miscalibrated distributions. We identify that miscalibration can be attributed to the use of sample-based metrics such as $\operatorname{minMPJPE}$. In a series of simulations, we show that minimizing $\operatorname{minMPJPE}$, as commonly done, should converge to the correct mean prediction. However, it fails to correctly capture the uncertainty, thus resulting in a miscalibrated distribution. To mitigate this problem, we propose an accurate and well-calibrated model called Conditional Graph Normalizing Flow (cGNFs). Our model is structured such that a single cGNF can estimate both conditional and marginal densities within the same model - effectively solving a zero-shot density estimation problem. We evaluate cGNF on the Human 3.6M dataset and show that cGNF provides a well-calibrated distribution estimate while being close to state-of-the-art in terms of overall $\operatorname{minMPJPE}$. Furthermore, cGNF outperforms previous methods on occluded joints while it remains well-calibrated.

****

Learning to Abstain from Uninformative Data

Yikai Zhang,Songzhu Zheng,Mina Dalirrooyfard,Pengxiang Wu,Anderson Schneider,Yuriy Nevmyvaka,Chao Chen

Learning and decision making in domains with naturally high noise-to-signal ratios – such as Finance or Healthcare – can be challenging yet extremely important. In this paper, we study a problem of learning and decision making under a general noisy generative process. The distribution has a significant proportion of uninformative data with high noise in label, while part of the data contains useful information represented by low label noise. This dichotomy is present during both training and inference, which requires the proper handling of uninformative data at testing time. We propose a novel approach to learn under these conditions via a loss inspired by the selective learning theory. By minimizing the loss, the model is guaranteed to make a near-optimal decision by distinguishing informative data from the uninformative data and making predictions. We build upon the strength of our theoretical guarantees by describing an iterative algorithm, which jointly optimizes both a predictor and a selector, and evaluate its empirical performance under a variety of settings.

****

Order Matters: Agent-by-agent Policy Optimization

Xihuai Wang,Zheng Tian,Ziyu Wan,Ying Wen,Jun Wang,Weinan Zhang

While multi-agent trust region algorithms have achieved great success empirically in solving coordination tasks, most of them, however, suffer from a non-stationarity problem since agents update their policies simultaneously. In contrast, a sequential scheme that updates policies agent-by-agent provides another perspective and shows strong performance. However, sample inefficiency and lack of monotonic improvement guarantees for each agent are still the two significant challenges for the sequential scheme. In this paper, we propose the \textbf{A}gent-by-\textbf{a}gent \textbf{P}olicy \textbf{O}ptimization (A2PO) algorithm to improve the sample efficiency and retain the guarantees of monotonic improvement for each agent during training. We justify the tightness of the monotonic improvement bound compared with other trust region algorithms. From the perspective of sequentially updating agents, we further consider the effect of agent updating order and extend the theory of non-stationarity into the sequential update scheme. To evaluate A2PO, we conduct a comprehensive empirical study on four benchmarks: StarCraftII, Multi-agent MuJoCo, Multi-agent Particle Environment, and Google Research Football full game scenarios. A2PO consistently outperforms strong baselines.

****

AQuaMaM: An Autoregressive, Quaternion Manifold Model for Rapidly Estimating Complex SO(3) Distributions

Michael A. Alcorn

Accurately modeling complex, multimodal distributions is necessary for optimal decision-making, but doing so for rotations in three-dimensions, i.e., the SO(3) group, is challenging due to the curvature of the rotation manifold. The recently described implicit-PDF (IPDF) is a simple, elegant, and effective approach for learning arbitrary distributions on SO(3) up to a given precision. However, inference with IPDF requires N forward passes through the network's final multilayer perceptron—where N places an upper bound on the likelihood that can be calculated by the model—which is prohibitively slow for those without the computational resources necessary to parallelize the queries. In this paper, I introduce AQuaMaM, a neural network capable of both learning complex distributions on the rotation manifold and calculating exact likelihoods for query rotations in a single forward pass. Specifically, AQuaMaM autoregressively models the projected components of unit quaternions as a mixture of uniform distributions that partition their geometrically-restricted domain of values. On an "infinite" toy dataset with ambiguous viewpoints, AQuaMaM rapidly converges to a sampling distribution closely matching the true data distribution. In contrast, the sampling distribution for IPDF dramatically diverges from the true data distribution, despite IPDF approaching its theoretical minimum evaluation loss during training. On a constructed dataset of 500,000 renders of a die in different rotations, an AQuaMaM model trained from scratch reaches a log-likelihood 14% higher than an IPDF model using a pretrained ResNet-50. Further, compared to IPDF, AQuaMaM uses 24% fewer parameters, has a prediction throughput 52$\times$ faster on a single GPU, and converges in a similar amount of time during training.

****

Conformal Prediction is Robust to Label Noise

Bat-Sheva Einbinder,Stephen Bates,Anastasios Nikolas Angelopoulos,Asaf Gendler,Yaniv Romano

We study the robustness of conformal prediction—a powerful tool for uncertainty quantification—to label noise. Our analysis tackles both regression and classification problems, characterizing when and how it is possible to construct uncertainty sets that correctly cover the unobserved noiseless ground truth labels. Through stylized theoretical examples and practical experiments, we argue that naïve conformal prediction covers the noiseless ground truth label unless the noise distribution is adversarially designed. This leads us to believe that correcting for label noise is unnecessary except for pathological data distributions or noise sources. In such cases, we can also correct for noise of bounded size in the conformal prediction algorithm in order to ensure correct coverage of the ground truth labels without score or data regularity.

****

Φ-DVAE: Learning Physically Interpretable Representations with Nonlinear Filtering

Alex John Glyn-Davies,Connor Duffin,Omer Deniz Akyildiz,Mark Girolami

Incorporating unstructured data into physical models is a challenging problem that is emerging in data assimilation. Traditional approaches focus on well-defined observation operators whose functional forms are typically assumed to be known. This prevents these methods from achieving a consistent model-data synthesis in configurations where the mapping from data-space to model-space is unknown. To address these shortcomings, in this paper we develop a physics-informed dynamical variational autoencoder (Φ-DVAE) for embedding diverse data streams into time-evolving physical systems described by differential equations. Our approach combines a standard (possibly nonlinear) filter for the latent state-space mod el and a VAE, to embed the unstructured data stream into the latent dynamical system. A variational Bayesian framework is used for the joint estimation of the embedding, latent states, and unknown system parameters. To demonstrate the method, we look at three examples: video datasets generated by the advection and Korteweg-de Vries partial differential equations, and a velocity field generated by the Lorenz-63 system. Comparisons with relevant baselines show that the Φ-DVAE provides a data efficient dynamics encoding methodology that is competitive with standard approaches, with the added benefit of incorporating a physically interpretable latent space.

****

Revisiting Structured Dropout

Yiren Zhao,Oluwatomisin Dada,Xitong Gao,Robert D. Mullins

Large neural networks are often overparameterised and prone to overfitting, Dropout is a widely used regularization technique to combat overfitting and improve model generalization. However, unstructured Dropout is not always effective for specific network architectures and this has led to the formation of multiple structured Dropout approaches to improve model performance and, sometimes, reduce the computational resources required for inferencing. In this work we revisit structured Dropout comparing different Dropout approaches on natural language processing and computer vision tasks for multiple state-of-the-art networks. Additionally, we devise an approach to structured Dropout we call \textbf{\emph{ProbDropBlock}} which drops contiguous blocks from feature maps with a probability given by the normalized feature salience values. We find that with a simple scheduling strategy the proposed approach to structured Dropout consistently improved model performance compared to baselines and other Dropout approaches on a diverse range of tasks and models. In particular, we show \textbf{\emph{ProbDropBlock}} improves RoBERTa finetuning on MNLI by 0.22%, and training of ResNet50 on ImageNet by 0.28%.

****

Flatter, Faster: Scaling Momentum for Optimal Speedup of SGD

Aditya Cowsik,Tankut Can,Paolo Glorioso

Commonly used optimization algorithms often show a trade-off between good generalization and fast training times. For instance, stochastic gradient descent (SGD) tends to have good generalization; however, adaptive gradient methods have superior training times. Momentum can help accelerate training with SGD, but so far there has been no principled way to select the momentum hyperparameter. Here we study implicit bias arising from the interplay between SGD with label noise and momentum in the training of overparameterized neural networks. We find that scaling the momentum hyperparameter $1-\beta$ with the learning rate to the power of $2/3$ maximally accelerates training, without sacrificing generalization. To analytically derive this result we develop an architecture-independent framework, where the main assumption is the existence of a degenerate manifold of global minimizers, as is natural in overparameterized models. Training dynamics display the emergence of two characteristic timescales that are well-separated for generic values of the hyperparameters. The maximum acceleration of training is reached when these two timescales meet, which in turn determines the scaling limit we propose. We perform experiments, including matrix sensing and ResNet on CIFAR10, which provide evidence for the robustness of these results.

****

Learning implicit hidden Markov models using neural likelihood-free inference

Sanmitra Ghosh,Paul Birrell,Daniela De Angelis

Likelihood-free inference methods for implicit models based on neural conditional density estimation were shown to drastically reduce the simulation burden in comparison to classical methods such as ABC. However, when applied in the context of any latent variable model, such as a Hidden Markov model (HMM), these methods are designed to only estimate the parameters rather than the joint posterior distribution of both the parameters and the hidden states. Naive application of these methods to a HMM, ignoring the inference of this joint posterior distribution, will result in overestimation of uncertainty of the posterior predictive. We propose a postprocessing step that can rectify this problem. Our approach relie s on learning directly the intractable posterior distribution of the hidden states, using an autoregressive-flow, by exploiting the Markov property. Upon evaluating our approach on some intractable HMMs, we found that the quality of the estimates retrieved using our postprocessing is comparable to what can be achieved using a computationally expensive particle-filtering which additionally requires a tractable data distribution.

****

Brain Signal Generation and Data Augmentation with a Single-Step Diffusion Probabilistic Model

Szabolcs Torma,Dr. Luca Szegletes

Brain-computer interfaces based on deep learning rely on large amounts of high-quality data. Finding publicly available brain signal datasets that meet all requirements is a challenge. However, brain signals synthesized with generative models may provide a solution to this problem. Our work builds on diffusion probabilistic models (DPMs) and aims to generate brain signals that have the properties needed to develop further classification models based on deep learning. We show that our DPM can generate high-quality event-related potentials (ERPs) and motor imagery (MI) signals. Furthermore, with the progressive distillation of the model, subject-specific data can be produced in a one-step reverse process. We augment publicly available datasets and demonstrate the impact of the generated signals on a deep learning classification model. DPMs are versatile models, and this work shows that brain signal processing is one of many other tasks in which these models can be useful.

****

Know Your Boundaries: The Advantage of Explicit Behavior Cloning in Offline RL

Wonjoon Goo,Scott Niekum

We introduce an offline reinforcement learning (RL) algorithm that explicitly clones a behavior policy to constrain value learning. In offline RL, it is often important to prevent a policy from selecting unobserved actions, since the consequence of these actions cannot be presumed without additional information about the environment. One straightforward way to implement such a constraint is to explicitly model a given data distribution via behavior cloning and directly force a policy not to select uncertain actions. However, many offline RL methods instantiate the constraint indirectly---for example, pessimistic value estimation---due to a concern about errors when modeling a potentially complex behavior policy. In this work, we argue that it is not only viable but beneficial to explicitly model the behavior policy for offline RL because the constraint can be realized in a stable way with the trained model. We first suggest a theoretical framework that allows us to incorporate behavior-cloned models into value-based offline RL methods, enjoying the strength of both explicit behavior cloning and value learning. Then, we propose a practical method utilizing a score-based generative model for behavior cloning. With the proposed method, we show state-of-the-art performance on several datasets within the D4RL and Robomimic benchmarks and achieve competitive performance across all datasets tested.

****

On the Convergence of AdaGrad(Norm) on $\mathbb{R}^d$: Beyond Convexity, Non-Asymptotic Rate and Acceleration

Zijian Liu,Ta Duy Nguyen,Alina Ene,Huy Nguyen

Existing analysis of AdaGrad and other adaptive methods for smooth convex optimization is typically for functions with bounded domain diameter. In unconstrained problems, previous works guarantee an asymptotic convergence rate without an explicit constant factor that holds true for the entire function class. Furthermore, in the stochastic setting, only a modified version of AdaGrad, different from the one commonly used in practice, in which the latest gradient is not used to update the stepsize, has been analyzed. Our paper aims at bridging these gaps and developing a deeper understanding of AdaGrad and its variants in the standard setting of smooth convex functions as well as the more general setting of quasar convex functions. First, we demonstrate new techniques to explicitly bound the convergence rate of the vanilla AdaGrad for unconstrained problems in both deterministic and stochastic settings. Second, we propose a variant of AdaGrad for wh ich we can show the convergence of the last iterate, instead of the average iterate. Finally, we give new accelerated adaptive algorithms and their convergence guarantee in the deterministic setting with explicit dependency on the problem parameters, improving upon the asymptotic rate shown in previous works.

****

Bounded Attacks and Robustness in Image Transform Domains

Mohamed-Hicham LEGHETTAS,Markus Püschel

Classical image transformation such as the discrete cosine transform (DCT) and the discrete wavelet transforms (DWTs) provide semantically meaningful representations of images. In this paper we propose a general method for adversarial attacks in such transform domains that, in contrast to prior work, obey the L^∞ constraint in the pixel domain. The key idea is to replace the standard attack based on projections with the barrier method. Experiments with DCT and DWTs produce adversarial examples that are significantly more similar to the original than with prior attacks. Further, through adversarial training we show that robustness against our attacks transfers to robustness against a broad class of common image perturbations.

****

SP2 : A Second Order Stochastic Polyak Method

Shuang Li,William Joseph Swartworth,Martin Taká■,Deanna Needell,Robert M. Gower

Recently the SP (Stochastic Polyak step size) method has emerged as a competitive adaptive method for setting the step sizes of SGD. SP can be interpreted as a method specialized to interpolated models, since it solves the interpolation equations. SP solves these equation by using local linearizations of the model. We take a step further and develop a method for solving the interpolation equations that uses the local second-order approximation of the model. Our resulting method SP2 uses Hessian-vector products to speed-up the convergence of SP. Furthermore, and rather uniquely among second-order methods, the design of SP2 in no way relies on positive definite Hessian matrices or convexity of the objective function. We show SP2 is competitive both in experiments and in theory.

We show SP2 is very competitive on matrix completion, non-convex test problems and logistic regression. We also provide a convergence theory on sums-of-quadratics.

****

Multi-Objective GFlowNets

Moksh Jain,Sharath Chandra Raparthy,Alex Hernández-García,Jarrid Rector-Brooks,Yoshua Bengio,Santiago Miret,Emmanuel Bengio

In many applications of machine learning, like drug discovery and material design, the goal is to generate candidates that simultaneously maximize a set of objectives. As these objectives are often conflicting, there is no single candidate that simultaneously maximizes all objectives, but rather a set of Pareto-optimal candidates where one objective cannot be improved without worsening another. Moreover, these objectives, when considered in practice are often under-specified, making diversity of candidates a key consideration. The existing multi-objective optimization methods focus predominantly on covering the Pareto front, failing the capture diversity in the space of candidates. Motivated by the success of GFlowNets for generation of diverse candidates in a single objective setting, in this paper we consider Multi-Objective GFlowNets (MOGFNs). MOGFNs consist of a Conditional GFlowNet which models a family of single-objective sub-problems derived by decomposing the multi-objective optimization problem. Our work is the first to empirically demonstrate conditional GFlowNets. Through a series of experiments on synthetic tasks and real-world domains, we empirically demonstrate that MOGFNs outperform existing methods in terms of Hypervolume, R2-distance and candidate diversity. We also demonstrate the effectiveness of MOGFNs over existing methods in active learning settings. Finally, we supplement our empirical results with a careful analysis of each component of MOGFNs.

****

Making Better Decision by Directly Planning in Continuous Control

Jinhua Zhu,Yue Wang,Lijun Wu,Tao Qin,Wengang Zhou,Tie-Yan Liu,Houqiang Li

By properly utilizing the learned environment model, model-based reinforcement l earning methods can improve the sample efficiency for decision-making problems. Beyond using the learned environment model to train a policy, the success of MCTS-based methods shows that directly incorporating the learned environment model as a planner to make decisions might be more effective. However, when action space is of high dimension and continuous, directly planning according to the learned model is costly and non-trivial. Because of two challenges: (1) the infinite number of candidate actions and (2) the temporal dependency between actions in different timesteps. To address these challenges, inspired by Differential Dynamic Programming (DDP) in optimal control theory, we design a novel Policy Optimization with Model Planning (POMP) algorithm, which incorporates a carefully designed Deep Differential Dynamic Programming (D3P) planner into the model-based RL framework. In D3P planner, (1) to effectively plan in the continuous action space, we construct a locally quadratic programming problem that uses a gradient-based optimization process to replace search. (2) To take the temporal dependency of actions at different timesteps into account, we leverage the updated and latest actions of previous timesteps (i.e., step $1, \cdots, h-1$) to update the action of the current step (i.e., step h), instead of updating all actions simultaneously. We theoretically prove the convergence rate for our D3P planner and analyze the effect of the feedback term. In practice, to effectively apply the neural network based D3P planner in reinforcement learning, we leverage the policy network to initialize the action sequence and keep the action update conservative in the planning process. Experiments demonstrate that POMP consistently improves sample efficiency on widely used continuous control tasks. Our code is released at https://github.com/POMP-D3P/POMP-D3P.

****

Large language models are not zero-shot communicators

Laura Eline Ruis,Akbir Khan,Stella Biderman,Sara Hooker,Tim Rocktäschel,Edward Grefenstette

The recent success of large language models (LLMs) has drawn heavy attention and investment in their use as conversational and embodied systems. Despite widespread use of LLMs as conversational agents, evaluations of performance fail to capture a crucial aspect of communication: interpreting language in context. Humans interpret language using beliefs, prior knowledge about the world, and more. For example, we intuitively understand the response "I wore gloves" to the question "Did you leave fingerprints?" as meaning "No". To investigate whether LLMs have the ability to make this type of inference, known as an implicature, we design a simple task and evaluate a set of models. We find that despite only evaluating on utterances that require a binary inference (yes or no), most perform close to random. Models adapted to be "aligned with human intent" via reinforcement learning perform much better, but still leave a significant gap with human performance. This gap is even more pronounced for context-heavy utterances. We present our findings as the starting gun for further research into evaluating how LLMs interpret language in context, in order to drive the development of more pragmatic and useful models of human discourse.

****

Data dependent frequency sensitivity of convolutional neural networks

Charles Godfrey,Elise Bishoff,Myles McKay,Davis Brown,Grayson Jorgenson,Henry Kvinge,Eleanor Byler

It is widely acknowledged that trained convolutional neural networks (CNNs) have different levels of sensitivity to signals of different frequency. In particular, a number of empirical studies have documented CNNs sensitivity to low-frequency signals. In this work we show with theory and experiments that this observed sensitivity is a consequence of the frequency distribution of natural images, which is known to have most of its power concentrated in low-to-mid frequencies. Our theoretical analysis relies on representations of the layers of a CNN in frequency space, an idea that has previously been used to accelerate computations and study implicit bias of network training algorithms, but to the best of our knowledge has not been applied in the domain of model robustness.

****

Is end-to-end learning enough for fitness activity recognition?

Antoine Mercier,Guillaume Berger,Sunny Panchal,Florian Dietrichkeit,Cornelius Böhm,Ingo Bax,Roland Memisevic

End-to-end learning has taken hold of many computer vision tasks, in particular, related to still images, with task-specific optimization yielding very strong performance. Nevertheless, human-centric action recognition is still largely dominated by hand-crafted pipelines, and only individual components are replaced by neural networks that typically operate on individual frames. As a testbed to study the relevance of such pipelines, we present a new fully annotated video dataset of fitness activities. Any recognition capabilities in this domain are almost exclusively a function of human poses and their temporal dynamics, so pose-based solutions should perform well. We show that, with this labelled data, end-to-end learning on raw pixels can compete with state-of-the-art action recognition pipelines based on pose estimation. We also show that end-to-end learning can support temporally fine-grained tasks such as real-time repetition counting.

****

Efficient Exploration using Model-Based Quality-Diversity with Gradients

Bryan Lim,Manon Flageat,Antoine Cully

Exploration is a key challenge in Reinforcement Learning, especially in long-horizon, deceptive and sparse-reward environments. For such applications, population-based approaches have proven effective. Methods such as Quality-Diversity deals with this by encouraging novel solutions and producing a diversity of behaviours. However, these methods are driven by either undirected sampling (i.e. mutations) or use approximated gradients (i.e. Evolution Strategies) in the parameter space, which makes them highly sample-inefficient. In this paper, we propose a model-based Quality-Diversity approach, relying on gradients and learning in imagination. Our approach optimizes all members of a population simultaneously to maintain both performance and diversity efficiently by leveraging the effectiveness of QD algorithms as good data generators to train deep models. We demonstrate that it maintains the divergent search capabilities of population-based approaches while significantly improving their sample efficiency (5 times faster) and quality of solutions (2 times more performant).

****

HiT-MDP: Learning the SMDP option framework on MDPs with Hidden Temporal Embeddings

Chang Li,Dongjin Song,Dacheng Tao

The standard option framework is developed on the Semi-Markov Decision Process (SMDP) which is unstable to optimize and sample inefficient. To this end, we propose the Hidden Temporal MDP (HiT-MDP) and prove that the option-induced HiT-MDP is homomorphic equivalent to the option-induced SMDP. A novel transformer-based framework is introduced to learn options' embedding vectors (rather than conventional option tuples) on HiT-MDPs. We then derive a stable and sample efficient option discovering method under the maximum-entropy policy gradient framework. Extensive experiments on challenging Mujoco environments demonstrate HiT-MDP's efficiency and effectiveness: under widely used configurations, HiT-MDP achieves competitive, if not better, performance compared to the state-of-the-art baselines on all finite horizon and transfer learning environments. Moreover, HiT-MDP significantly outperforms all baselines on infinite horizon environments while exhibiting smaller variance, faster convergence, and better interpretability. Our work potentially sheds light on the theoretical ground of extending the option framework into a large-scale foundation model.

****

Improved Group Robustness via Classifier Retraining on Independent Splits

Thien Hang Nguyen,Hongyang Ryan Zhang,Huy Nguyen

Deep neural networks learned by minimizing the average risk can achieve strong average performance, but their performance for a subgroup may degrade, if the subgroup is underrepresented in the overall data population. Group distributionally robust optimization (Sagawa et al., 2020a, GDRO) is a standard baseline for learning models with strong worst-group performance. However, GDRO requires group labels for every example during training and can be prone to overfitting, often requiring careful model capacity control via regularization or early stopping. Wh en only a limited amount of group labels is available, Just Train Twice (Liu et al., 2021, JTT) is a popular approach which infers a pseudo-group-label for every unlabeled example. The process of inferring pseudo labels can be highly sensitive during model selection. To alleviate overfitting for GDRO and the pseudo labeling process for JTT, we propose a new method via classifier retraining on independent splits (of the training data). We find that using a novel sample splitting procedure achieves robust worst-group performance in the fine-tuning step. When evaluated on benchmark image and text classification tasks, our approach consistently reduces the requirement of group labels and hyperparameter search during training. Experimental results confirm that our approach performs favorably compared with existing methods (including GDRO and JTT) when either group labels are available during training or are only available during validation.

****

(Certified!!) Adversarial Robustness for Free!

Nicholas Carlini,Florian Tramer,Krishnamurthy Dj Dvijotham,Leslie Rice,Mingjie Sun,J Zico Kolter

In this paper we show how to achieve state-of-the-art certified adversarial robustness to 2-norm bounded perturbations by relying exclusively on off-the-shelf pretrained models. To do so, we instantiate the denoised smoothing approach of Salman et al. by combining a pretrained denoising diffusion probabilistic model and a standard high-accuracy classifier. This allows us to certify 71% accuracy on ImageNet under adversarial perturbations constrained to be within a 2-norm of 0.5, an improvement of 14 percentage points over the prior certified SoTA using any approach, or an improvement of 30 percentage points over denoised smoothing. We obtain these results using only pretrained diffusion models and image classifiers, without requiring any fine tuning or retraining of model parameters.

****

URVoice: An Akl-Toussaint/ Graham- Sklansky Approach towards Convex Hull Computation for Sign Language Interpretation

Madhumitha V,Santhi Natarajan,Bharathi Malarkeddy A

We present URVoice, a vocalizer for the communication impaired, based on the Indian Sign Language Notations. Contemporary psychological theories consider language and speech as devices to understand complex psychological processes and deliver them as cultural products of ideas and communication. Sign and gesture language, offering an intelligent co-ordination of eye-and-hand and ear-and-mouth, has evolved as an intelligent manifestation of speech for the impaired. However, they have very limited modality and iconicity in accommodating a greater range of linguistically relevant meanings. URVoice is an Augmentative and Alternative Communication (AAC) device, which currently features a pipeline of forward communication from signer to collocutor with a novel approach shouldered on convex hull using vision-based approach. The solution achieves real time translation of gesture to text/voice using convex hull as the computational geometry, which follows Akl-Toussaint heuristic and Graham-Sklansky scan algorithms. The results are weighed against our other solutions based on conventional Machine Learning and Deep Learning approaches. A futuristic version of URVoice, with voice translated to sign language gestures, will be a complete solution for effectively bridging the cognitive and communication gap between the impaired and the abled lot.

****

Gaussian-Bernoulli RBMs Without Tears

Renjie Liao,Simon Kornblith,Mengye Ren,David J. Fleet,Geoffrey Hinton

We revisit the challenging problem of training Gaussian-Bernoulli restricted Boltzmann machines (GRBMs), introducing two innovations. We propose a novel Gibbs-Langevin sampling algorithm that outperforms existing methods like Gibbs sampling.

We propose a modified contrastive divergence (CD) algorithm so that one can generate images with GRBMs starting from noise.

This enables direct comparison of GRBMs with deep generative models, improving evaluation protocols in the RBM literature.

Moreover, we show that modified CD and gradient clipping are enough to robustly train GRBMs with large learning rates, thus removing the necessity of various tr

icks in the literature.
Experiments on Gaussian Mixtures, MNIST, FashionMNIST, and CelebA show GRBMs can generate good samples, despite their single-hidden-layer architecture.
****

Efficient Conditionally Invariant Representation Learning

Roman Pogodin,Namrata Deka,Yazhe Li,Danica J. Sutherland,Victor Veitch,Arthur Gretton

We introduce the Conditional Independence Regression CovariancE (CIRCE), a measure of conditional independence for multivariate continuous-valued variables. CIRCE applies as a regularizer in settings where we wish to learn neural features $\varphi(X)$ of data X to estimate a target Y, while being conditionally independent of a distractor Z given Y. Both Z and Y are assumed to be continuous-valued but relatively low dimensional, whereas X and its features may be complex and high dimensional. Relevant settings include domain-invariant learning, fairness, and causal learning. The procedure requires just a single ridge regression from Y to kernelized features of Z, which can be done in advance. It is then only necessary to enforce independence of $\varphi(X)$ from residuals of this regression, which is possible with attractive estimation properties and consistency guarantees. By contrast, earlier measures of conditional feature dependence require multiple regressions for each step of feature learning, resulting in more severe bias and variance, and greater computational cost. When sufficiently rich features are used, we establish that CIRCE is zero if and only if $\varphi(X) \perp \!\!\! \perp Z \mid Y$. In experiments, we show superior performance to previous methods on challenging benchmarks, including learning conditionally invariant image features. Code for image data experiments is available at github.com/namratadeka/circe.
****

Heterogeneous Neuronal and Synaptic Dynamics for Spike-Efficient Unsupervised Learning: Theory and Design Principles

Biswadeep Chakraborty,Saibal Mukhopadhyay

This paper shows that the heterogeneity in neuronal and synaptic dynamics reduces the spiking activity of a Recurrent Spiking Neural Network (RSNN) while improving prediction performance, enabling spike-efficient (unsupervised) learning.
We analytically show that the diversity in neurons' integration/relaxation dynamics improves an RSNN's ability to learn more distinct input patterns (higher memory capacity), leading to improved classification and prediction performance. We further prove that heterogeneous Spike-Timing-Dependent-Plasticity (STDP) dynamics of synapses reduce spiking activity but preserve memory capacity. The analytical results motivate Heterogeneous RSNN design using Bayesian optimization to determine heterogeneity in neurons and synapses to improve $\mathcal{E}$, defined as the ratio of spiking activity and memory capacity. The empirical results on time series classification and prediction tasks show that optimized HRSNN increases performance and reduces spiking activity compared to a homogeneous RSNN.
****

MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises

Emanuele Palumbo,Imant Daunhawer,Julia E Vogt

Multimodal VAEs have recently gained attention as efficient models for weakly-supervised generative learning with multiple modalities. However, all existing variants of multimodal VAEs are affected by a non-trivial trade-off between generative quality and generative coherence. In particular mixture-based models achieve good coherence only at the expense of sample diversity and a resulting lack of generative quality. We present a novel variant of the mixture-of-experts multimodal variational autoencoder that improves its generative quality, while maintaining high semantic coherence. We model shared and modality-specific information in separate latent subspaces, proposing an objective that overcomes certain dependencies on hyperparameters that arise for existing approaches with the same latent space structure. Compared to these existing approaches, we show increased robustness with respect to changes in the design of the latent space, in terms of the capacity allocated to modality-specific subspaces. We show that our model achieves both good generative coherence and high generative quality in challenging experiments, including more complex multimodal datasets than those used in previous works.

****

A probabilistic framework for task-aligned intra- and inter-area neural manifold estimation

Edoardo Balzani,Jean-Paul G Noel,Pedro Herrero-Vidal,Dora E Angelaki,Cristina Savin

Latent manifolds provide a compact characterization of neural population activity and of shared co-variability across brain areas. Nonetheless, existing statistical tools for extracting neural manifolds face limitations in terms of interpretability of latents with respect to task variables, and can be hard to apply to datasets with no trial repeats. Here we propose a novel probabilistic framework that allows for interpretable partitioning of population variability within and across areas in the context of naturalistic behavior. Our approach for task aligned manifold estimation (TAME-GP) explicitly partitions variability into private and shared sources which can themselves be subdivided in task-relevant and task irrelevant components, uses a realistic Poisson noise model, and introduces temporal smoothing of latent trajectories in the form of a Gaussian Process prior. This TAME-GP graphical model allows for robust estimation of task-relevant variability in local population responses, and of shared co-variability between brain areas. We demonstrate the efficiency of our estimator on within model and biologically motivated simulated data. We also apply it to several datasets of neural population recordings during behavior. Overall, our results demonstrate the capacity of TAME-GP to capture meaningful intra- and inter-area neural variability with single trial resolution.

****

Applying Second Order Optimization to Deep Transformers with Parameter-Efficient Tuning

Ning Ding,Qiaosen Wang,Yulin Chen,Pengjun Xie,Zhiyuan Liu,Hai-Tao Zheng,Maosong Sun

Despite the theoretical superiority in convergence issues, second-order optimizers are generally not among the top choices for training large-scale neural networks due to their high computational and memory cost. Nevertheless, introduced in recent progress of parameter-efficient tuning is a new paradigm that large-scale pre-trained models (PTMs) can be adapted to specific tasks by optimizing a tiny proportion of parameters, which might hopefully change the game. We associate this new paradigm with the computational tractability of second-order optimizers and succeed in applying them to large PTMs that are from hundreds of millions to billions in scale. Beyond verifying their tractability, we further investigate the stability-influencing factors in the optimization process and propose accordingly a Newton-step-clipping approach in which we clip the update tensors rather than the gradients. This approach stabilizes the convergence by gating the magnitude of Newton steps along the optimization trajectories through the rugged landscapes of deep transformers.

We conduct extensive experiments across different downstream tasks, demonstrating that, when equipped with our Newton-step-clipping strategy, second-order optimizers, especially Kronecker-factored curvature approximation (K-FAC), can attain comparable and even superior results and faster convergence to those state-of-the-art bars implemented with AdamW. Furthermore, we scale the model up to 3 billion parameters and validate the tractability and effectiveness of our method. This work is not only the first successful application of second-order optimization on such large-scale models but also sheds light on the possibility of further optimization-wise analysis on large-scale models in the future.

****

Density Sketches for Sampling and Estimation

Aditya Desai,Benjamin Coleman,Anshumali Shrivastava

There has been an exponential increase in the data generated worldwide. Insights into this data led by machine learning (ML) have given rise to exciting applications such as recommendation engines, conversational agents, and so on. Often, data for these applications is generated at a rate faster than ML pipelines can c

onsume it. In this paper, we propose Density Sketches(DS) - a cheap and practical approach to reducing data redundancy in a streaming fashion. DS creates a succinct online summary of data distribution. While DS does not store the samples from the stream, we can sample unseen data on the fly from DS to use for downstream learning tasks. In this sense, DS can replace actual data in many machine learning pipelines analogous to generative models. Importantly, unlike generative models, which do not have statistical guarantees, the sampling distribution of DS asymptotically converges to underlying unknown density distribution.
****

Mask-tuning: Towards Improving Pre-trained Language Models' Generalization
Somayeh Ghanbarzadeh,Hamid Palangi,Yan Huang,Radames Cruz Moreno,Hamed Khanpour

Pre-trained language models have the known generalization problem. This issue emerges from the pre-trained language models' learning process that heavily relies on spurious correlations, which work for the majority of training examples but do not hold in general. As a consequence, the models' performance drops substantially on out-of-distribution datasets. Previous studies proposed various solutions, including data augmentation and learning process improvement. In this paper, we present Mask-tuning, an approach that alleviates the impact of spurious correlations on the fine-tuning learning process. To achieve this goal, Mask-tuning integrates masked language training into the fine-tuning learning process. In this case, Mask-tuning perturbs the linguistic relation of downstream tasks' training examples and computes masked language training loss. Then, the perturbed examples are fed into fine-tuning process to be classified based on their ground-truth label and compute the fine-tuning training loss. Afterward, Mask-tuning loss -- a weighted aggregation of masked language model training loss and fine-tuning loss-- updates the masked language model and fine-tuning through training iterations. Extensive experiments show that Mask-tuning consistently improves the pre-trained language models' generalization on out-of-distribution datasets and enhances their performance on in-distribution datasets. The source code and pre-trained models will be available on the author's GitHub page.
****

Meta-Learning via Classifier(-free) Guidance
Elvis Nava,Seijin Kobayashi,Yifei Yin,Robert K. Katzschmann,Benjamin F Grewe

State-of-the-art meta-learning techniques do not optimize for zero-shot adaptation to unseen tasks, a setting in which humans excel. On the contrary, meta-learning algorithms learn hyperparameters and weight initializations that explicitly optimize for few-shot learning performance. In this work, we take inspiration from recent advances in generative modeling and language-conditioned image synthesis to propose meta-learning techniques that use natural language guidance to achieve higher zero-shot performance compared to the state-of-the-art. We do so by recasting the meta-learning problem as a multi-modal generative modeling problem: given a task, we consider its adapted neural network weights and its natural language description as equivalent multi-modal task representations. We first train an unconditional generative hypernetwork model to produce neural network weights; then we train a second "guidance" model that, given a natural language task description, traverses the hypernetwork latent space to find high-performance task-adapted weights in a zero-shot manner. We explore two alternative approaches for latent space guidance: "HyperCLIP"-based classifier guidance and a conditional Hypernetwork Latent Diffusion Model ("HyperLDM"), which we show to benefit from the classifier-free guidance technique common in image generation. Finally, we demonstrate that our approaches outperform existing meta-learning methods with zero-shot learning experiments on our Meta-VQA dataset, which we specifically constructed to reflect the multi-modal meta-learning setting.
****

Tiered Pruning for Efficient Differentialble Inference-Aware Neural Architecture Search
Slawomir Kierat,Mateusz Sieniawski,Denys Fridman,Chenhan D. Yu,Szymon Migacz,Pawel Morkisz,Alex Fit-Florea

We propose three novel pruning techniques to improve the cost and results of inference-aware Differentiable Neural Architecture Search (DNAS). First, we introdu ce $\textbf{Prunode}$, a stochastic bi-path building block for DNAS, which can search over inner hidden dimensions with $\mathcal{O}(1)$ memory and compute complexity. Second, we present an algorithm for pruning blocks within a stochastic layer of the SuperNet during the search. Third, we describe a novel technique for pruning unnecessary stochastic layers during the search. The optimized models resulting from the search are called PruNet and establishes a new state-of-the-art Pareto frontier for NVIDIA V100 in terms of inference latency for ImageNet Top-1 image classification accuracy. PruNet as a backbone also outperforms GPUNet and EfficientNet on the COCO object detection task on inference latency relative to mean Average Precision (mAP).

****

MyoDex: Generalizable Representations for Dexterous Physiological Manipulation
Vittorio Caggiano,Sudeep Dasari,Vikash Kumar

The complexity of human dexterity has attracted attention from multiple fields. Still, much is to be understood about how hand manipulation behaviors emerge. In this work we aim at learning dexterous manipulation behaviors with a physiologically realistic hand model: MyoHand. In contrast to prior works demonstrating isolated postural and force control, here we demonstrate musculoskeletal agents (MyoDex) exhibiting contact-rich dynamic dexterous manipulation behaviors in simulation. Furthermore, to demonstrate generalization, we show that a single MyoDex agent can be trained to solve up-to 14 different contact-rich tasks. Aligned with human development, simultaneous learning of multiple tasks imparts physiological coordinated muscle contractions i.e., muscle synergies, that are not only shared amongst those in-domain tasks but are also effective in out-of-domain tasks. By leveraging these pre-trained manipulation synergies, we show generalization to 14 additional previously unsolved tasks. While physiological behaviors with large muscle groups (such as legged-locomotion, arm-reaching, etc), have been demonstrated before, to the best of our knowledge nimble behaviors of this complexity with smaller muscle groups are being demonstrated for the first time.

****

Do We Really Need Labels for Backdoor Defense?
Zidi Xiong,Dongxian Wu,Yifei Wang,Yisen Wang

Since training a model from scratch always requires massive computational resources recently, it has become popular to download pre-trained backbones from third-party platforms and deploy them in various downstream tasks. While providing some convenience, it also introduces potential security risks like backdoor attacks, which lead to target misclassification for any input image with a specifically defined trigger (i.e., backdoored examples). Current backdoor defense methods always rely on clean labeled data, which indicates that safely deploying the pre-trained model in downstream tasks still demands these costly or hard-to-obtain labels. In this paper, we focus on how to purify a backdoored backbone with only unlabeled data. To evoke the backdoor patterns without labels, we propose to leverage the unsupervised contrastive loss to search for backdoors in the feature space. Surprisingly, we find that we can mimic backdoored examples with adversarial examples crafted by contrastive loss, and erase them with adversarial finetuning. Thus, we name our method as Contrastive Backdoor Defense (CBD). Against several backdoored backbones from both supervised and self-supervised learning, extensive experiments demonstrate our unsupervised method achieves comparable or even better defense compared to these supervised backdoor defense methods. Thus, our method allows practitioners to safely deploy pre-trained backbones on downstream tasks without extra labeling costs.

****

Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics
Shoaib Ahmed Siddiqui,Nitarshan Rajkumar,Tegan Maharaj,David Krueger,Sara Hooker

Modern machine learning research relies on relatively few carefully curated data sets. Even in these datasets, and typically in `untidy' or raw data, practitioners are faced with significant issues of data quality and diversity which can be prohibitively labor intensive to address. Existing methods for dealing with these challenges tend to make strong assumptions about the particular issues at play, and often require a priori knowledge or metadata such as domain labels. Our wo rk is orthogonal to these methods: we instead focus on providing a unified and efficient framework for Metadata Archaeology -- uncovering and inferring metadata of examples in a dataset. We curate different subsets of data that might exist in a dataset (e.g. mislabeled, atypical, or out-of-distribution examples) using simple transformations, and leverage differences in learning dynamics between these probe suites to infer metadata of interest. Our method is on par with far more sophisticated mitigation methods across different tasks: identifying and correcting mislabeled examples, classifying minority-group samples, prioritizing points relevant for training and enabling scalable human auditing of relevant examples.

****

Single SMPC Invocation DPHelmet: Differentially Private Distributed Learning on a Large Scale

Moritz Kirschte,Sebastian Meiser,Saman Ardalan,Esfandiar Mohammadi

Distributing machine learning predictors enables the collection of large-scale datasets while leaving sensitive raw data at trustworthy sites. We introduce a learning technique that is scalable to a large number of users, satisfies Differential Privacy, and is applicable to non-trivial tasks, such as CIFAR-10. For a large number of participants, communication cost is one of the main challenges. We achieve a low communication cost by requiring only a single invocation of an efficient secure multiparty summation protocol. By relying on state-of-the-art feature extractors, we are able to utilize differentially private convex learners for non-trivial tasks such as CIFAR-10. Convex learners have proven to have a strong utility-private tradeoff. Our experimental results show that for $1{,}000$ users with 50 data points each, our scheme outperforms state-of-the-art scalable distributed learning methods (differentially private federated learning, short DP-FL) while requiring around 500 times fewer communication costs: For CIFAR-10, we achieve a classification accuracy of $67.3\,\%$ for an $\varepsilon = 0.59$ while DP-FL achieves $57.6\,\%$. We also show the learnability properties convergence and uniform stability.

****

Re-Benchmarking Out-of-Distribution Detection in Deep Neural Networks

Jinling Gao,Haoyue Bai,Lin Zhu,Nanyang Ye

Out-of-distribution (OOD) detection is a key challenge for making machine learning models robust in the real world, where we want models to be aware of uncertainty outside their training data distribution. Despite the rapid development of existing OOD detection algorithms, their experimental settings are usually inconsistent, e.g., datasets, evaluation metrics, model selection, implementation choices. In this paper, we aim to understand OOD detection fundamentally and provide a comprehensive benchmarking of the current state of the art OOD detection methods in a consistent and realistic evaluation setting. This benchmarking contains a serious of datasets split, model selection criteria and OOD detection algorithms. This experimental framework can be easily extended to new algorithms, datasets, and model selection criteria. We conduct extensive experiments on this benchmark and find that the threshold of OOD detection algorithms are not consistent over different datasets and model selection criteria.

****

Towards Antisymmetric Neural Ansatz Separation

Aaron Zweig,Joan Bruna

We study separations between two fundamental models (or \emph{Ansätze}) of antisymmetric functions, that is, functions f of the form $f(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) = \text{sign}(\sigma)f(x_1, \ldots, x_N)$, where σ is any permutation.
These arise in the context of quantum chemistry, and are the basic modeling tool for wavefunctions of Fermionic systems.
Specifically, we consider two popular antisymmetric Ansätze: the Slater representation, which leverages the alternating structure of determinants, and the Jastrow ansatz, which augments Slater determinants with a product by an arbitrary symmetric function. We construct an antisymmetric function that can be more efficiently expressed in Jastrow form, yet provably cannot be approximated by Slater de

terminants unless there are exponentially (in N^2) many terms. This represents the first explicit quantitative separation between these two Ansätze.
****

Multi-instance Interactive Segmentation with Self-Supervised Transformer
Xavier Jiménez,Jaonary Rabarisoa,Valentin Belissen,Quoc Cuong PHAM
The rise of Vision Transformers (ViT) combined with better self-supervised learning pre-tasks has taken representation learning to the next level, beating supervised results on ImageNet. In particular, self-attention mechanism of ViT allows to easily visualize semantic information learned by the network. Following revealing of attention maps of DINO, many tried to leverage its representations for unsupervised segmentation. Despite very promising results for basic images with a single clear object in a simple background, representation of ViT are not able to segment images, with several classes and object instance, in an unsupervised fashion yet. In this paper, we propose SALT: Semi-supervised Segmentation with Self-supervised Attention Layers in Transformers, an interactive algorithm for multi-class/multi-instance segmentation. We follow previous works path and take it a step further by discriminating between different objects, using sparse human help to select said objects. We show that remarkable results are achieved with very sparse labels. Different pre-tasks are compared, and we show that self-supervised ones are more robust for panoptic segmentation, and overall achieve very similar performance. Evaluation is carried out on Pascal VOC 2007 and COCO-panoptic. Performance is evaluated for extreme conditions such as very noisy, and sparse interactions going to as little as one interaction per class.
****

Spurious Features in Continual Learning
Timothee LESORT
Continual Learning (CL) is the research field addressing learning without forgetting when the data distribution is not static.

This paper studies spurious features' influence on continual learning algorithms.

We show that continual learning algorithms solve tasks by selecting features that are not generalizable.

Our experiments highlight that continual learning algorithms face two related problems: (1) spurious features (SP) and (2) local spurious features (LSP). The first one is due to a covariate shift between training and testing data, while the second is due to the limited access to data at each training step.

We study (1) through a consistent set of continual learning experiments varying spurious correlation amount and data distribution support.

We show that (2) is a major cause of performance decrease in continual learning along with catastrophic forgetting.

This paper presents a different way of understanding performance decrease in continual learning by highlighting the influence of (local) spurious features in algorithms capabilities.
****

Time Series Subsequence Anomaly Detection via Graph Neural Networks
Weiqi Chen,Zhiqiang Zhou,Qingsong Wen,Liang Sun
Time series subsequence anomaly detection is an important task in a large variety of real-world applications ranging from health monitoring to AIOps, and is challenging due to complicated underlying temporal dynamics and unpredictable anomalous patterns. Firstly, how to effectively learn the temporal dependency in time series remains a challenge. Secondly, diverse and complicated anomalous subsequences as well as the lack of labels make accurate detection difficult. For example, the popular subsequence anomaly detection algorithm---time series discord---fails to handle recurring anomalies. Thirdly, many existing algorithms require a proper subsequence length for effective detection, which is difficult or impossible in practice. In this paper, we present a novel approach to subsequence anomaly detection which combines practical heuristics of time series discords and temporal relationships with deep neural networks. By performing length selection considering multi-scale information and incorporating prior knowledge using graph neural networks, our method can adaptively learn the appropriate subsequence l

ength as well as integrated representations from both priors and raw data favorable to anomaly detection. In particular, our graph incorporates both semantic and temporal relationships between subsequences. The experimental results demonstrate the effectiveness of the proposed algorithm, which achieves superior performance on multiple time series anomaly benchmarks in comparison with state-of-the-art algorithms.

****

Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness

Joel Dapello,Kohitij Kar,Martin Schrimpf,Robert Baldwin Geary,Michael Ferguson,David Daniel Cox,James J. DiCarlo

While some state-of-the-art artificial neural network systems in computer vision are strikingly accurate models of the corresponding primate visual processing, there are still many discrepancies between these models and the behavior of primates on object recognition tasks. Many current models suffer from extreme sensitivity to adversarial attacks and often do not align well with the image-by-image behavioral error patterns observed in humans. Previous research has provided strong evidence that primate object recognition behavior can be very accurately predicted by neural population activity in the inferior temporal (IT) cortex, a brain area in the late stages of the visual processing hierarchy. Therefore, here we directly test whether making the late stage representations of models more similar to that of macaque IT produces new models that exhibit more robust, primate-like behavior. We conducted chronic, large-scale multi-electrode recordings across the IT cortex in six non-human primates (rhesus macaques). We then use these data to fine-tune (end-to-end) the model "IT" representations such that they are more aligned with the biological IT representations, while preserving accuracy on object recognition tasks. We generate a cohort of models with a range of IT similarity scores validated on held-out animals across two image sets with distinct statistics. Across a battery of optimization conditions, we observed a strong correlation between the models' IT-likeness and alignment with human behavior, as well as an increase in its adversarial robustness. We further assessed the limitations of this approach and find that the improvements in behavioral alignment and adversarial robustness generalize across different image statistics, but not to object categories outside of those covered in our IT training set. Taken together, our results demonstrate that building models that are more aligned with the primate brain leads to more robust and human-like behavior, and call for larger neural data-sets to further augment these gains.

****

On the Expressive Power of Geometric Graph Neural Networks

Chaitanya K. Joshi,Cristian Bodnar,Simon V Mathis,Taco Cohen,Pietro Lio

The expressive power of Graph Neural Networks (GNNs) has been studied extensively through the lens of the Weisfeiler-Leman (WL) graph isomorphism test. Yet, many graphs arising in real-world applications come embedded in Euclidean space with an additional notion of geometric isomorphism, which is not covered by the WL framework. In this work, we propose a geometric version of the WL test (GWL) for discriminating geometric graphs while respecting the underlying physical symmetries: permutation, rotation, reflection, and translation. We use GWL to characterise the expressive power of GNNs that are invariant or equivariant to physical symmetries by studying the classes of geometric graphs that can or cannot be distinguished by these architectures. This allows us to formalise the advantages equivariant GNN layers have over their invariant counterparts in the Geometric Deep Learning blueprint. Finally, we connect our discrimination-based perspective with the universal approximation properties of geometric GNNs and prove they are two sides of the same coin.

****

Fusion over the Grassmann Manifold for Incomplete-Data Clustering

Jeremy Scott Johnson,Huanran Li,Chun Gan,Zheng Lu,Matthew Malloy,Daniel L. Pimentel-Alarcón

This paper presents a new paradigm to cluster incomplete vectors using subspaces as proxies to exploit the geometry of the Grassmannian. We leverage this new pe

rspective to develop an algorithm to cluster and complete data in a union of subspaces via a fusion penalty formulation. Our approach does not require prior knowledge of the number of subspaces, is naturally suited to handle noise, and only requires an upper bound on the subspaces’ dimensions. In developing our model, we present local convergence guarantees. We describe clustering, completion, model selection, and sketching techniques that can be used in practice, and complement our analysis with synthetic and real-data experiments.
****

Off Policy Average Reward Actor Critic with Deterministic Policy Search
Naman Saxena,Subhojyoti Khastagir,Shishir N Y,Shalabh Bhatnagar
The average reward criterion is relatively less explored as most existing works in the Reinforcement Learning literature consider the discounted reward criterion. There are few recent works that present on-policy average reward actor-critic algorithms, but average reward off-policy actor-critic is relatively less explored. In this paper, we present both on-policy and off-policy deterministic policy gradient theorems for the average reward performance criterion. Using these theorems, we also present an Average Reward Off-Policy Deep Deterministic Policy Gradient (ARO-DDPG) Algorithm. We show a finite time analysis of the resulting three-timescale stochastic approximation scheme and obtain an ϵ-optimal stationary policy with a sample complexity of $\Omega(\epsilon^{-2.5})$. We compare the average reward performance of our proposed algorithm and observe better empirical performance compared to state-of-the-art on-policy average reward actor-critic algorithms over MuJoCo based environments.
****

Why Did This Model Forecast This Future? Information-Theoretic Temporal Saliency for Counterfactual Explanations of Probabilistic Forecasts
Chirag Raman,Hayley Hung,Marco Loog
Probabilistic forecasting of multivariate time series is significant to several research domains where multiple futures exist for a single observed sequence. Identifying the observations on which a well-performing model bases its forecasts can enable domain experts to form data-driven hypotheses about the causal relationships between features. Consequently, we begin by revisiting the question: what constitutes a causal explanation? One hurdle in the landscape of explainable artificial intelligence is that what constitutes an explanation is not well-grounded. We build upon Miller's framework of explanations derived from research in multiple social science disciplines, and establish a conceptual link between counterfactual reasoning and saliency-based explanation techniques. However, the complication is a lack of a consistent and principled notion of saliency. Also, commonly derived saliency maps may be inconsistent with the data generation process and the underlying model. We therefore leverage a unifying definition of information-theoretic saliency grounded in preattentive human visual cognition and extend it to forecasting settings. In contrast to existing methods that require either explicit training of the saliency mechanism or access to the internal parameters of the underlying model, we obtain a closed-form solution for the resulting saliency map for commonly used density functions in probabilistic forecasting. To empirically evaluate our explainability framework in a principled manner, we construct a synthetic dataset of conversation dynamics and demonstrate that our method recovers the true salient timesteps for a forecast given a well-performing underlying model.
****

CLMIU: Commonsense Learning in Multimodal Image Understanding.
Sergio Sanchez Santiesteban,Muhammad Awais,Sara Atito,Yi-Zhe Song,Josef Kittler
The problem of automatically describing the content of an image through accurate and meaningful captions has been attracting considerable attention among computer vision researchers. Recently, Transformers have been applied to image captioning to encode cross-modal information, in conjunction with Convolutional Neural Networks, which provide image region descriptions in terms of embeddings and object labels as input. However, the generated captions sometimes fail to capture the intentions, relationships, and abstract concepts that rely on general or commonsense knowledge. In this work we propose a novel network design, combining the

strengths of Transformer models with graph-based models conveying external (common sense) knowledge. Our proposed architecture is a pure vision transformer-based image captioning model, with sequences of image patches used directly as input, without extracting any regional features. In particular, unlike the prior work, our architecture incorporates a knowledge-augmented encoder with a Transformer backbone to inject the external knowledge extracted from a knowledge graph. Furthermore, the bidirectional training on a vision-language corpus of image-text pairs, using modality specific self-supervised learning objectives, achieves promising results compared to the state-of-the-art. Our method has been trained from scratch on a small dataset, achieving a 3.8%, 2.7%, 3.2% and 6.3% improvement in BLEU@4, Meteor, Rouge and Cider scores respectively. We also reported competitive results on the NoCaps dataset, showing that the model generalizes to unseen object categories.

****

In-Situ Text-Only Adaptation of Speech Models with Low-Overhead Speech Imputations

Ashish Mittal,Sunita Sarawagi,Preethi Jyothi

Fast and accurate adaptation of automatic speech recognition (ASR) systems using only text data in the target domain is a problem of long-standing practical relevance. Text-only adaptation was easy in traditional cascaded ASR systems with completely decoupled acoustic and language models. Recently, the RNNTransducer (RNN-T) has emerged as a default ASR model because of its high accuracy, low latency, and capability of supporting streaming input. However text-only adaptation of the RNN-T model is significantly more challenging due to its tight integration of acoustic and language models and end-to-end training. Existing recent approaches for text-only adaptation of RNN-Ts, either entail significant modification to the network or introduce high latency during decoding. We propose a new approach (TOLSTOI) that imputes speech representations internal to a baseline RNN-T, starting from text-only inputs, and performs in-situ adaptation that results in higher adaptation accuracy without any runtime overheads during decoding. Our imputation model is a function of the labeled data and trained parameters of the ASR model, and that we show, is more effective in controlling catastrophic forgetting compared to existing methods. We establish the effectiveness of TOLSTOI using three target domains and two ASR models of varying complexity. We yield up to 35% relative reduction in word error rate with text-only adaptation while forgetting the least compared to existing adaptation approaches. Our method is easy to implement and can be harnessed on existing RNN-T models without requiring ASR model training from scratch.

****

Rethinking Uniformity in Self-Supervised Representation Learning

Xianghong Fang,Jian Li,Xiangchu Feng,Benyou Wang

Self-supervised representation learning has achieved great success in many machine learning tasks. While many research efforts focus on learning better representations by preventing the model from the \emph{collapse} problem, less attention has been drawn to analyzing the collapse degrees of representations. In this paper, we present a formal study of collapse analysis via the \emph{uniformity} metric, which measures how uniformly learned representations distribute on the surface of the unit hypersphere. We fundamentally find that \textit{representation that obeys zero-mean isotropic Gaussian distribution is with the ideal uniformity} since its l_2-normalized form uniformly distributes on the surface of the unit hypersphere. Therefore, we propose to use the Wasserstein distance between the distribution of learned representations and the ideal distribution as a quantifiable metric of \emph{uniformity}. Moreover, we design five desirable constraints for ideal uniformity metrics, based on which we find that the proposed uniformity metric satisfies all constraints while the existing one does not. Synthetic experiments also demonstrate the proposed uniformity metric is capable to deal with the dimensional collapse while the existing one is insensitive. Furthermore, we impose the proposed \emph{uniformity} metric as an auxiliary loss term for various existing self-supervised methods, which consistently improves the downstream performance.

****

Proposal-Contrastive Pretraining for Object Detection from Fewer Data

Quentin Bouniot,Romaric Audigier,Angelique Loesch,Amaury Habrard

The use of pretrained deep neural networks represents an attractive way to achieve strong results with few data available. When specialized in dense problems such as object detection, learning local rather than global information in images has proven to be more efficient. However, for unsupervised pretraining, the popular contrastive learning requires a large batch size and, therefore, a lot of resources. To address this problem, we are interested in transformer-based object detectors that have recently gained traction in the community with good performance and with the particularity of generating many diverse object proposals.

In this work, we present Proposal Selection Contrast (ProSeCo), a novel unsupervised overall pretraining approach that leverages this property. ProSeCo uses the large number of object proposals generated by the detector for contrastive learning, which allows the use of a smaller batch size, combined with object-level features to learn local information in the images. To improve the effectiveness of the contrastive loss, we introduce the object location information in the selection of positive examples to take into account multiple overlapping object proposals. When reusing pretrained backbone, we advocate for consistency in learning local information between the backbone and the detection head.

We show that our method outperforms state of the art in unsupervised pretraining for object detection on standard and novel benchmarks in learning with fewer data.

****

SkillS: Adaptive Skill Sequencing for Efficient Temporally-Extended Exploration

Giulia Vezzani,Dhruva Tirumala,Markus Wulfmeier,Dushyant Rao,Abbas Abdolmaleki,Ben Moran,Tuomas Haarnoja,Jan Humplik,Roland Hafner,Michael Neunert,Claudio Fantacci,Tim Hertweck,Thomas Lampe,Fereshteh Sadeghi,Nicolas Heess,Martin Riedmiller

The ability to effectively reuse prior knowledge is a key requirement when building general and flexible Reinforcement Learning (RL) agents.

Skill reuse is one of the most common approaches, but current methods have considerable limitations. For example, fine-tuning an existing policy frequently fails, as the policy can degrade rapidly early in training, particularly in sparse reward tasks. In a similar vein, distillation of expert behavior can lead to poor results when given sub-optimal experts.

We compare several common approaches for skill transfer on multiple domains and in several different transfer settings, including under changes in task and system dynamics. We identify how existing methods can fail and introduce an alternative approach which sidesteps some of these problems.

Our approach learns to sequence existing temporally-abstract skills for exploration but learns the final policy directly from the raw experience. This conceptual split enables rapid adaptation and thus efficient data collection but without constraining the final solution. Our approach significantly outperforms many classical methods across a suite of evaluation tasks and we use a broad set of ablations to highlight the importance of different components of our method.

****

Bridging between Pool- and Stream-Based Active Learning with Temporal Data Coherence

Sebastian Schmidt,Stephan Günnemann

Active learning (AL) reduces the amount of labeled data needed for training a machine learning model by choosing intelligently which instances to label. Classic pool-based AL needs all data to be present in a datacenter, which can be challenging with the increasing amounts of data needed in deep learning. However, AL on mobile devices and robots like autonomous cars can filter the data from perception sensor streams before it ever reaches the datacenter. In our work, we investigate AL for such image streams and propose a new concept exploiting their temporal properties. We define three methods using a pseudo uncertainty based on loss learning (Yoo & Kweon, 2019). The first considers the temporal change of uncertainty and requires 5% less labeled data than the vanilla approach. It is

extended by the change in latent space in the second method. The third method, temporal distance loss stream (TDLS), combines both with submodular optimization . In our evaluation on an extension of the public Audi Autonomous Driving Dataset (Geyer et al., 2020) we outperform state-of-the-art approaches by using 1% fewer labels. Additionally, we compare our stream-based approaches with existing approaches for AL in a pool-based scenario. Our experiments show that, although pool-based AL has access to more data, our stream-based AL approaches need 0.5% fewer labels.
****

Scaling Laws For Deep Learning Based Image Reconstruction

Tobit Klug,Reinhard Heckel

Deep neural networks trained end-to-end to map a measurement of a (noisy) image to a clean image perform excellent for a variety of linear inverse problems.
Current methods are only trained on a few hundreds or thousands of images as opposed to the millions of examples deep networks are trained on in other domains.
In this work, we study whether major performance gains are expected from scaling up the training set size.
We consider image denoising, accelerated magnetic resonance imaging, and super-resolution and empirically determine the reconstruction quality as a function of training set size, while simultaneously scaling the network size.
For all three tasks we find that an initially steep power-law scaling slows significantly already at moderate training set sizes.
Interpolating those scaling laws suggests that even training on millions of images would not significantly improve performance.
To understand the expected behavior, we analytically characterize the performance of a linear estimator learned with early stopped gradient descent.
The result formalizes the intuition that once the error induced by learning the signal model is small relative to the error floor, more training examples do not improve performance.
****

Robust Exploration via Clustering-based Online Density Estimation

Alaa Saade,Steven Kapturowski,Daniele Calandriello,Charles Blundell,Michal Valko,Pablo Sprechmann,Bilal Piot

Intrinsic motivation is a critical ingredient in reinforcement learning to enable progress when rewards are sparse. However, many existing approaches that measure the novelty of observations are brittle, or rely on restrictive assumptions about the environment which limit generality. We propose to decompose the exploration problem into two orthogonal sub-problems: (i) finding the right representation (metric) for exploration (ii) estimating densities in this representation space.

To address (ii), we introduce Robust Exploration via Clustering-based Online Density Estimation (RECODE), a non-parametric method that estimates visitation counts for clusters of states that are similar according to the metric induced by any arbitrary representation learning technique. We adapt classical clustering algorithms to design a new type of memory that allows RECODE to keep track of the history of interactions over thousands of episodes, thus effectively tracking global visitation counts. This is in contrast to existing non-parametric approaches, that can only store the recent history, typically the current episode.

The generality of RECODE allows us to easily address (i) by leveraging both off-the-shelf and novel representation learning techniques. In particular, we introduce a novel generalization of the action-prediction representation that leverages multi-step predictions and that we find to be better suited to a suite of challenging 3D-exploration tasks in DM-HARD-8. We show experimentally that our approach can work with a variety of RL agents, and obtain state-of-the-art performance on Atari and DM-HARD-8.
****

Meta Learning to Bridge Vision and Language Models for Multimodal Few-Shot Learning

Ivona Najdenkoska,Xiantong Zhen,Marcel Worring

Multimodal few-shot learning is challenging due to the large domain gap between vision and language modalities. Existing methods are trying to communicate visual concepts as prompts to frozen language models, but rely on hand-engineered task induction to reduce the hypothesis space. To make the whole process learnable, we introduce a multimodal meta-learning approach. Specifically, our approach decomposes the training of the model into a set of related multimodal few-shot tasks. We define a meta-mapper network, acting as a meta-learner, to efficiently bridge frozen large-scale vision and language models and leverage their already learned capacity. By updating the learnable parameters only of the meta-mapper, it learns to accrue shared meta-knowledge among these tasks. Thus, it can rapidly adapt to newly presented samples with only a few gradient updates. Importantly, it induces the task in a completely data-driven manner, with no need for a hand-engineered task induction. We evaluate our approach on recently proposed multimodal few-shot benchmarks, measuring how rapidly the model can bind novel visual concepts to words and answer visual questions by observing only a limited set of labeled examples. The experimental results show that our meta-learning approach outperforms the baseline across multiple datasets and various training settings while being computationally more efficient.

****

DLP: Data-Driven Label-Poisoning Backdoor Attack

Xun Xian,Xuan Bi,Mingyi Hong,Jie Ding

Backdoor attacks, which aim to disrupt or paralyze classifiers on specific tasks, are becoming an emerging concern in several learning scenarios, e.g., Machine Learning as a Service (MLaaS). Various backdoor attacks have been introduced in the literature, including perturbation-based methods, which modify a subset of training data; and clean-sample methods, which relabel only a proportion of training samples. Indeed, clean-sample attacks can be particularly stealthy since they never require modifying the samples at the training and test stages. However, the state-of-the-art clean-sample attack of relabelling training data based on their semantic meanings could be ineffective and inefficient in test performances due to heuristic selections of semantic patterns. In this work, we introduce a new type of clean-sample backdoor attack, named as DLP backdoor attack, allowing attackers to backdoor effectively, as measured by test performances, for an arbitrary backdoor sample size. The critical component of DLP is a data-driven backdoor scoring mechanism embedding in a multi-task formulation, which enables attackers to simultaneously perform well on the normal learning tasks and the backdoor tasks. Systematic empirical evaluations show the superior performance of the proposed DLP to state-of-the-art clean-sample attacks.

****

AlphaFold Distillation for Improved Inverse Protein Folding

Igor Melnyk,Aurelie Lozano,Payel Das,Vijil Chenthamarakshan

Inverse protein folding, i.e., designing sequences that fold into a given three-dimensional structure, is one of the fundamental design challenges in bio-engineering and drug discovery. Traditionally, inverse folding mainly involves learning from sequences that have an experimentally resolved structure. However, the known structures cover only a tiny space of the protein sequences, imposing limitations on the model learning. Recently proposed forward folding models, e.g., AlphaFold, offer unprecedented opportunity for accurate estimation of the structure given a protein sequence. Naturally, incorporating a forward folding model as a component of an inverse folding approach offers the potential of significantly improving the inverse folding, as the folding model can provide a feedback on any generated sequence in the form of the predicted protein structure or a structural confidence metric. However, at present, these forward folding models are still prohibitively slow to be a part of the model optimization loop during training. In this work, we propose to perform knowledge distillation on the folding model's confidence metrics, e.g., pTM or pLDDT scores, to obtain a smaller, faster and end-to-end differentiable distilled model, which then can be included as part of the structure consistency regularized inverse folding model training. Moreover, our regularization technique is general enough and can be applied in oth er design tasks, e.g., sequence-based protein infilling. Extensive experiments show a clear benefit of our method over the non-regularized baselines. E.g., in inverse folding design problems we observe up to 3% improvement in sequence recovery and up to 45% improvement in protein diversity, while still preserving structural consistency of the generated sequences.

****

Convexifying Transformers: Improving optimization and understanding of transformer networks

Tolga Ergen,Behnam Neyshabur,Harsh Mehta

Understanding the fundamental mechanism behind the success of transformer networks is still an open problem in the deep learning literature. Although their remarkable performance has been mostly attributed to the self-attention mechanism, the literature still lacks a solid analysis of these networks and interpretation of the functions learned by them. To this end, we study the training problem of attention/transformer networks and introduce a novel convex analytic approach to improve the understanding and optimization of these networks. Particularly, we first introduce a convex alternative to the self-attention mechanism and reformulate the regularized training problem of attention/transformer networks. Then, we cast the reformulation as a convex optimization problem that is interpretable and easier to optimize. Moreover, as a byproduct of our convex analysis, we reveal an implicit regularization mechanism, which promotes sparsity across tokens. Therefore, we not only improve the optimization of attention/transformer networks but also provide a solid theoretical understanding of the functions learned by them. We also demonstrate the effectiveness of our theory through several numerical experiments.

****

Unsupervised Model-based Pre-training for Data-efficient Control from Pixels

Sai Rajeswar,Pietro Mazzaglia,Tim Verbelen,Alexandre Piché,Bart Dhoedt,Aaron Courville,Alexandre Lacoste

Controlling artificial agents from visual sensory data is an arduous task. Reinforcement learning (RL) algorithms can succeed in this but require large amounts of interactions between the agent and the environment. To alleviate the issue, unsupervised RL proposes to employ self-supervised interaction and learning, for adapting faster to future tasks. Yet, whether current unsupervised strategies improve generalization capabilities is still unclear, especially in visual control settings. In this work, we design an unsupervised RL strategy for data-efficient visual control. First, we show that world models pre-trained with data collected using unsupervised RL can facilitate adaptation for future tasks. Then, we analyze several design choices to adapt faster, effectively reusing the agents' pre-trained components, and planning in imagination, with our hybrid planner, which we dub Dyna-MPC. By combining the findings of a large-scale empirical study, we establish an approach that strongly improves performance on the Unsupervised RL Benchmark, requiring 20$\times$ less data to match the performance of supervised methods. The approach also demonstrates robust performance on the Real-Word RL benchmark, hinting that the approach generalizes to noisy environments.

****

A Cognitive-inspired Multi-Module Architecture for Continual Learning

Shruthi Gowda,Bahram Zonooz,Elahe Arani

Artificial neural networks (ANNs) exhibit a narrow scope of expertise on stationary independent data. However, data in the real world is continuous and dynamic, and ANNs must adapt to novel scenarios while also retaining the learned knowledge to become lifelong learners. The ability of humans to excel at these tasks can be attributed to multiple factors ranging from cognitive computational structures, cognitive biases, and the multi-memory systems in the brain. We incorporate key concepts from each of these to design a cognitive-inspired continual learning method. Cognitive Continual Learner (CCL) includes multiple modules, implicit and explicit knowledge representation dichotomy, inductive bias, and a multi-memory system. CCL shows improvement across all continual learning settings and also exhibits reduced task recency bias. To test versatility of continual learning methods on a challenging distribution shift, we introduce a novel domain-increm

ental dataset Domain^{2}IL. In addition to improved performance on existing benchmarks, CCL also demonstrates superior performance on this dataset.
****

Shuffled Transformers for Blind Training

Hengyuan Xu,Liyao Xiang,Hangyu Ye,Dixi Yao,Pengzhi Chu

Conventional split learning faces the challenge of preserving training data and model privacy as a part of the training is beyond the data owner's control. We tackle this problem by introducing blind training, i.e., training without being aware of the data or the model, realized by shuffled Transformers. This is attributed to our intriguing findings that the inputs and the model weights of the Transformer encoder blocks, the backbone of Transformer, can be shuffled without degrading the model performance. We not only have proven the shuffling invariance property in theory, but also designs a privacy-preserving split learning framework following the property, with little modification to the original Transformer architecture. We carry out verification of the properties through experiments, and also show our proposed framework successfully defends privacy attacks to split learning with superiority.
****

Non-Gaussian Process Regression

Yaman Kindap,Simon J. Godsill

Standard GPs offer a flexible modelling tool for well-behaved processes. However, deviations from Gaussianity are expected to appear in real world datasets, with structural outliers and shocks routinely observed. In these cases GPs can fail to model uncertainty adequately and may over-smooth inferences. Here we extend the GP framework into a new class of time-changed GPs that allow for straightforward modelling of heavy-tailed non-Gaussian behaviours, while retaining a tractable conditional GP structure through an infinite mixture of non-homogeneous GPs representation. The conditional GP structure is obtained by conditioning the observations on a latent transformed input space and the random evolution of the latent transformation is modelled using a Lévy process which allows Bayesian inference in both the posterior predictive density and the latent transformation function. We present Markov chain Monte Carlo inference procedures for this model and demonstrate the potential benefits compared to a standard GP.
****

ImageNet-X: Understanding Model Mistakes with Factor of Variation Annotations

Badr Youbi Idrissi,Diane Bouchacourt,Randall Balestriero,Ivan Evtimov,Caner Hazirbas,Nicolas Ballas,Pascal Vincent,Michal Drozdzal,David Lopez-Paz,Mark Ibrahim

Deep learning vision systems are widely deployed across applications where reliability is critical. However, even today's best models can fail to recognize an object when its pose, lighting, or background varies. While existing benchmarks surface examples challenging for models, they do not explain why such mistakes arise. To address this need, we introduce ImageNet-X—a set of sixteen human annotations of factors such as pose, background, or lighting the entire ImageNet-1k validation set as well as a random subset of 12k training images. Equipped with ImageNet-X, we investigate 2,200 current recognition models and study the types of mistakes as a function of model’s (1) architecture, e.g. transformer vs. convolutional, (2) learning paradigm, e.g. supervised vs. self-supervised, and (3) training procedures, e.g., data augmentation. Regardless of these choices, we find models have consistent failure modes across ImageNet-X categories. We also find that while data augmentation can improve robustness to certain factors, they induce spill-over effects to other factors. For example, color-jitter augmentation improves robustness to color and brightness, but surprisingly hurts robustness to pose. Together, these insights suggest to advance the robustness of modern vision models, future research should focus on collecting additional data and understanding data augmentation schemes. Along with these insights, we release a toolkit based on ImageNet-X to spur further study into the mistakes image recognition systems make.
****

Hardware-aware compression with Random Operation Access Specific Tile (ROAST) hashing

Aditya Desai,Keren Zhou,Anshumali Shrivastava

Advancements in deep learning are often associated with increasing model sizes. Training and deploying large models require sophisticated hardware and incur significantly higher costs. Thus, model compression is a widely explored approach to solving the problem. However, SOTA techniques fall short in one or more desirable aspects of compression - for instance, pruning does not reduce memory for training, quantization can only provide up to $32\times$ compression, HashedNet is cache-inefficient, etc. This paper proposes a model-agnostic, cache-friendly, and hardware-aware model compression approach: Random Operation Access Specific Tile (ROAST) hashing. ROAST collapses the parameters by clubbing them through a lightweight mapping. While clubbing these parameters, ROAST utilizes cache hierarchies by aligning the memory access pattern with the parameter access pattern. ROAST is up to $\sim 25 \times$ faster to train and $\sim 50 \times$ faster to infer than the popular parameter sharing method HashedNet. Additionally, ROAST introduces global weight sharing, which is empirically and theoretically superior to local weight sharing in HashedNet, and can be of independent interest. With ROAST, we can efficiently train and deploy the model using a much smaller memory footprint ($\sim 10 \times - 100 \times$ lesser) in text and image classification tasks

****

SoftZoo: A Soft Robot Co-design Benchmark For Locomotion In Diverse Environments

Tsun-Hsuan Wang,Pingchuan Ma,Andrew Everett Spielberg,Zhou Xian,Hao Zhang,Joshua B. Tenenbaum,Daniela Rus,Chuang Gan

While significant research progress has been made in robot learning for control, unique challenges arise when simultaneously co-optimizing morphology. Existing work has typically been tailored for particular environments or representations. In order to more fully understand inherent design and performance tradeoffs and accelerate the development of new breeds of soft robots, a comprehensive virtual platform — with well-established tasks, environments, and evaluation metrics — is needed. In this work, we introduce SoftZoo, a soft robot co-design platform for locomotion in diverse environments. SoftZoo supports an extensive, naturally-inspired material set, including the ability to simulate environments such as flat ground, desert, wetland, clay, ice, snow, shallow water, and ocean. Further, it provides a variety of tasks relevant for soft robotics, including fast locomotion, agile turning, and path following, as well as differentiable design representations for morphology and control. Combined, these elements form a feature-rich platform for analysis and development of soft robot co-design algorithms. We benchmark prevalent representations and co-design algorithms, and shed light on 1) the interplay between environment, morphology, and behavior (2) the importance of design space representations 3) the ambiguity in muscle formation and controller synthesis and 4) the value of differentiable physics. We envision that SoftZoo will serve as a standard platform and template an approach toward the development of novel representations and algorithms for co-designing soft robots’ behavioral and morphological intelligence. Demos are available on our project page.

****

Smooth Mathematical Functions from Compact Neural Networks

INGI HONG

This is paper for the smooth function approximation by neural networks (NN). Mathematical or physical functions can be replaced by NN models through regression. In this study, we get NNs that generate highly accurate and highly smooth function, which only comprised of a few weight parameters, through discussing a few topics about regression. First, we reinterpret inside of NNs for regression; consequently, we propose a new activation function--integrated sigmoid linear unit (ISLU). Then special charateristics of metadata for regression, which is different from other data like image or sound, is discussed for improving the performance of neural networks. Finally, the one of a simple hierarchical NN that generate models substituting mathematical function is presented, and the new batch concept ``meta-batch" which improves the performance of NN several times more is introduced. The new activation function, meta-batch method, features of numerical data, meta-augmentation with metaparameters, and a structure of NN generating a compact multi-layer perceptron(MLP) are essential in this study.

****

Self-Supervised Learning of Maximum Manifold Capacity Representations

Thomas Edward Yerxa,Yilun Kuang,Eero P Simoncelli,SueYeon Chung

Self-supervised Learning (SSL) has recently emerged as a successful strategy for learning useful representations of images without relying on hand-assigned labels. Many such methods aim to learn a function that maps distinct views of the same scene or object to nearby points in the representation space. These methods are often justified by showing that they optimize an objective that is an approximation of (or correlated with) the mutual information between representations of different views. Here, we recast the problem from the perspective of manifold capacity, a recently developed measure for evaluating the quality of a representation. Specifically, we develop a contrastive learning framework that aims to maximize the number of linearly separable object manifolds, yielding a Maximum Manifold Capacity Representation (MMCR). We apply this method to unlabeled images, each augmented by a set of basic transformations, and find that it learns meaningful features using the standard linear evaluation protocol. Specifically, we find that MMCRs support performance on object recognition comparable to recently developed SSL frameworks, while providing more robustness to adversarial attacks. Finally, empirical analysis reveals the means by which compression of object manifolds gives rise to class separability.

****

TOWARDS AN OBJECTIVE EVALUATION OF THE TRUSTWORTHINESS OF CLASSIFIERS

Manish Chandra,Debasis Ganguly

With the widespread deployment of AI models in applications that impact human lives, research on model trustworthiness has become increasingly important, as a result of which model effectiveness alone (measured, e.g., with accuracy, F1, etc.) should not be the only criteria to evaluate predictive models; additionally the trustworthiness of these models should also be factored in. It has been argued that the features deemed important by a black-box model should be aligned with the human perception of the data, which in turn, should contribute to increasing the trustworthiness of a model. Existing research in XAI evaluates such alignments with user studies - the limitations being that these studies are subjective, difficult to reproduce, and consumes a large amount of time to conduct. We propose an evaluation framework, which provides a quantitative measure for trustworthiness of a black-box model, and hence, we are able to provide a fair comparison between a number of different black-box models. Our framework is applicable to both text and images, and our experiment results show that a model with a higher accuracy does not necessarily exhibit better trustworthiness.

****

Fine-grain Inference on Out-of-Distribution Data with Hierarchical Classification

Randolph Linderman,Jingyang Zhang,Nathan Inkawhich,Hai Li,Yiran Chen

Machine learning methods must be trusted to make appropriate decisions in real-world environments, even when faced with out-of-distribution (OOD) samples. Many current approaches simply aim to detect OOD examples and alert the user when an unrecognized input is given. However, when the OOD sample significantly overlaps with the training data, a binary anomaly detection is not interpretable or explainable, and provides little information to the user. We propose a new model for OOD detection that makes predictions at varying levels of granularity—as the inputs become more ambiguous, the model predictions become coarser and more conservative. Consider an animal classifier that encounters an unknown bird species and a car. Both cases are OOD, but the user gains more information if the classifier recognizes that its uncertainty over the particular species is too large and predicts “bird” instead of detecting it as OOD. Furthermore, we diagnose the classifier’s performance at each level of the hierarchy improving the explainabilit y and interpretability of the model’s predictions. We demonstrate the effectiveness of hierarchical classifiers for both fine- and coarse-grained OOD tasks.

****

ResGrad: Residual Denoising Diffusion Probabilistic Models for Text to Speech

Zehua Chen,Yihan Wu,Yichong Leng,Jiawei Chen,Haohe Liu,Xu Tan,Yang Cui,Ke Wang,Lei He,sheng zhao,Jiang Bian,Danilo Mandic

Denoising Diffusion Probabilistic Models (DDPMs) are emerging in text-to-speech (TTS) synthesis because of their strong capability of generating high-fidelity samples. However, their iterative refinement process in high-dimensional data space results in slow inference speed, which restricts their application in real-time systems. Previous works have explored speeding up by minimizing the number of inference steps but at the cost of sample quality. In this work, to improve the inference speed for DDPM-based TTS model while achieving high sample quality, we propose ResGrad, a lightweight diffusion model which learns to refine the output spectrogram of an existing TTS model (e.g., FastSpeech 2) by predicting the residual between the model output and the corresponding ground-truth speech. ResGrad has several advantages: 1) Compare with other acceleration methods for DDPM which need to synthesize speech from scratch, ResGrad reduces the complexity of task by changing the generation target from ground-truth mel-spectrogram to the residual, resulting into a more lightweight model and thus a smaller real-time factor. 2) ResGrad is employed in the inference process of the existing TTS model in a plug-and-play way, without re-training this model. We verify ResGrad on the single-speaker dataset LJSpeech and two more challenging datasets with multiple speakers (LibriTTS) and high sampling rate (VCTK). Experimental results show that in comparison with other speed-up methods of DDPMs: 1) ResGrad achieves better sample quality with the same inference speed measured by real-time factor; 2) with similar speech quality, ResGrad synthesizes speech faster than baseline methods by more than 10 times. Audio samples are available at \url{https://resgrad1.github.io/}.

****

The Adversarial Regulation of the Temporal Difference Loss Costs More Than Expected

Ezgi Korkmaz

Deep reinforcement learning research has enabled reaching significant performance levels for sequential decision making in MDPs with highly complex observations and state dynamics with the aid of deep neural networks. However, this aid came with a cost that is inherent to deep neural networks which have increased sensitivities towards indistinguishable peculiarly crafted non-robust directions. To alleviate these sensitivities several studies suggested techniques to cope with this problem via explicitly regulating the temporal difference loss for the worst-case sensitivity. In our study, we show that these worst-case regularization techniques come with a cost that intriguingly causes inconsistencies and overestimations in the state-action value functions. Furthermore, our results essentially demonstrate that vanilla trained deep reinforcement learning policies have more accurate and consistent estimates for the state-action values. We believe our results reveal foundational intrinsic properties of the adversarial training techniques and demonstrate the need to rethink the approach to robustness in deep reinforcement learning.

****

Beyond Link Prediction: On Pre-Training Knowledge Graph Embeddings

Daniel Ruffinelli,Rainer Gemulla

Knowledge graph embeddings (KGE) models provide low-dimensional representations of the entities and relations in a knowledge graph (KG). Most prior work focused on training and evaluating KGE models for the task of link prediction; the question of whether or not KGE models provide useful representations more generally remains largely open. In this work, we explore the suitability of KGE models (i) for more general graph-structure prediction tasks and (ii) for downstream tasks such as entity classification. For (i), we found that commonly trained KGE models often perform poorly at structural tasks other than link prediction. Based on this observation, we propose a more general multi-task training approach, which includes additional self-supervised tasks such as neighborhood prediction or domain prediction. In our experiments, these multi-task KGE models showed significantly better overall performance for structural prediction tasks. For (ii), we investigate whether KGE models provide useful features for a variety of downstream tasks. Here we view KGE models as a form of self-supervised pre-training and study the impact of both model training and model selection on downstream task performance. We found that multi-task pre-training can (but does not always) significantly improve performance and that KGE models can (but do not always) compete with or even outperform task-specific GNNs trained in a supervised fashion. Our work suggests that more research is needed on how to pre-train KGE models and on their suitability for downstream applications.

****

Masked Siamese ConvNets: Towards an Effective Masking Strategy for General-purpose Siamese Networks

Li Jing,Jiachen Zhu,Yann LeCun

Siamese Networks are a popular self-supervised learning framework that learns useful representation without human supervision by encouraging representations to be invariant to distortions. Existing methods heavily rely on hand-crafted augmentations, which are not easily adapted to new domains. To explore a general-purpose or domain-agnostic siamese network, we investigate using masking as augmentations in siamese networks. Recently, masking for siamese networks has only been shown useful with transformer architectures, e.g. MSN and data2vec. In this work, we identify the underlying problems of masking for siamese networks with arbitrary backbones, including ConvNets. We propose an effective and general-purpose masking strategy and demonstrate its effectiveness on various siamese network frameworks. Our method generally improves siamese networks' performances in the few-shot image classification, and object detection tasks.

****

Canary in a Coalmine: Better Membership Inference with Ensembled Adversarial Queries

Yuxin Wen,Arpit Bansal,Hamid Kazemi,Eitan Borgnia,Micah Goldblum,Jonas Geiping,Tom Goldstein

As industrial applications are increasingly automated by machine learning models, enforcing personal data ownership and intellectual property rights requires tracing training data back to their rightful owners. Membership inference algorithms approach this problem by using statistical techniques to discern whether a target sample was included in a model's training set. However, existing methods only utilize the unaltered target sample or simple augmentations of the target to compute statistics. Such a sparse sampling of the model's behavior carries little information, leading to poor inference capabilities. In this work, we use adversarial tools to directly optimize for queries that are discriminative and diverse. Our improvements achieve significantly more accurate membership inference than existing methods, especially in offline scenarios and in the low false-positive regime which is critical in legal settings.

****

Maximum Entropy Information Bottleneck for Confidence-aware Stochastic Embedding

Sungtae An,Nataraj Jammalamadaka,Eunji Chong

Stochastic embedding has several advantages over deterministic embedding, such as the capability of associating uncertainty with the resulting embedding and robustness to noisy data. This is especially useful when the input data has ambiguity (e.g., blurriness or corruption) which often happens with in-the-wild settings. Many existing methods for stochastic embedding are limited by the assumption that the embedding follows a standard normal distribution under the variational information bottleneck principle. We present a different variational approach to stochastic embedding in which maximum entropy acts as the bottleneck, which we call "Maximum Entropy Information Bottleneck" or MEIB. We show that models trained with the MEIB objective outperform existing methods in terms of regularization, perturbation robustness, probabilistic contrastive learning, and risk-controlled recognition performance.

****

Reprogramming Large Pretrained Language Models for Antibody Sequence Infilling
Igor Melnyk,Vijil Chenthamarakshan,Pin-Yu Chen,Payel Das,Amit Dhurandhar,Inkit Padhi,Devleena Das
Antibodies comprise the most versatile class of binding molecules, with numerous applications in biomedicine. Therapeutic antibody development requires designing novel and diverse sequences with improved properties, while maintaining the structural consistency. Computational design of antibodies involves unusual challenges relative to designing other classes of proteins, as antibodies comprise multiple long, variable, and unstructured loops at the complementarity-determining region (CDR) that determine the antigen binding affinity and specificity of an antibody. Recently, deep language models and graph neural networks have shown impressive success in antibody sequence generation. Since only a limited number of antibody structures are known, training a model using this limited data can lead to degraded performance, particularly lacking diversity in the generated samples. To address such issues, we leverage the method of Model Reprogramming (MR) here, which focuses on repurposing pretrained machine learning models for target domain tasks with scarce data, where it may be difficult to train a high-performing model from scratch. Prior works in MR have primarily focused on classification-based tasks. We extend the capabilities of reprogramming beyond classification tasks, and towards a more complex problem of antibody sequence generation. Specifically, we introduce Reprogramming for Protein Sequence Infilling, a framework in which pretrained natural language models are repurposed for protein sequence infilling via reprogramming, to infill protein sequence templates as a method of novel protein generation. For variable CDR sequence design, we formulate the task as text infilling that uses the constant region of an antibody as the sequence template. Results on antibody design benchmarks show that our reprogrammed model on low resourced antibody sequence dataset provides highly diverse CDR sequences, up to more than a two-fold increase of diversity over the baselines, without losing structural integrity and naturalness. The performance benefit of the reprogrammed model learning only from antibody sequences is more evident for longer CDR design or for multiple loop infilling at once, compared to existing graph-based models that require additional structural information. The generated sequences also demonstrate enhanced antigen binding specificity or virus neutralization ability.
****
Using semantic distance for diverse and sample efficient genetic programming
David Saxton,Chrisantha Fernando
Evolutionary methods, such as genetic programming, search a space of programs to find those with good fitness, often using mutations that manipulate the syntactic structure of programs without being aware of how they affect the semantics. For applications where the semantics are highly sensitive to small syntactic mutations, or where fitness evaluation is expensive, this can make learning programs intractable.

We introduce a mutation operator that yields mutated programs that are semantically far from previously evaluated programs, while still being semantically close to their parent. For function regression, this leads to an algorithm that is one to two orders of magnitude more sample efficient than other gradient-free methods, such as genetic programming, or learning the weights of a neural network using evolutionary strategies.

We show how this method can be applied to learning architecture-specific and general purpose neural network optimizers, and to reinforcement learning loss functions. The learnt components are simple, interpretable, high performance, and contain novel features not seen before such as weight growth.
****
Semi-parametric Prompt-Generation for Model Editing
Harshavardhan Kamarthi,Yen-Chang Hsu,Yilin Shen,Hongxia Jin
Large Language models are used in various downstream tasks with great success. However, changing specific knowledge or beliefs of a model (a.k.a. model editing)

efficiently to revise inaccurate predictions while not affecting all other cases is still challenging. Most previous methods compute gradients to change the model. These strategies generally work, paying the cost of high computing and memory complexity. The semi-parametric strategy has recently shown its effectiveness in alleviating the complexity via introducing memory to store the edits of knowledge. However, the memory does not have a proper mechanism to be utilized by a large pre-trained language model, limiting its generalizability to more complicated model editing scenarios. This work proposes a prompt generation mechanism to bridge
the gap. Our method encodes the edits as prefix prompts for language models, then has the large pre-trained language model perform inference with the prompts. In other words, the model is edited by prompts without changing model parameters. Our method, SEPROG, significantly outperforms state-of-art methods by up to 20% on entailed edit benchmarks and provides up to 30% better performance over gradient-based methods on non-entailed benchmarks. These advantages are achieved with much less computation and memory consumption, proving prompt generation’s great potential in model editing problems.

****

Improved Learning-augmented Algorithms for k-means and k-medians Clustering
Thy Dinh Nguyen,Anamay Chaturvedi,Huy Nguyen
We consider the problem of clustering in the learning-augmented setting. We are given a data set in d-dimensional Euclidean space, and a label for each data point given by a predictor indicating what subsets of points should be clustered together. This setting captures situations where we have access to some auxiliary information about the data set relevant for our clustering objective, for instance the labels output by a neural network. Following prior work, we assume that there are at most an $\alpha \in (0,c)$ for some $c<1$ fraction of false positives and false negatives in each predicted cluster, in the absence of which the labels would attain the optimal clustering cost OPT. For a dataset of size m, we propose a deterministic k-means algorithm that produces centers with an improved bound on the clustering cost compared to the previous randomized state-of-the-art algorithm while preserving the $O(d m \log m)$ runtime. Furthermore, our algorithm works even when the predictions are not very accurate, i.e., our cost bound holds for α up to $1/2$, an improvement from α being at most $1/7$ in previous work. For the k-medians problem we again improve upon prior work by achieving a biquadratic improvement in the dependence of the approximation factor on the accuracy parameter α to get a cost of $(1+O(\alpha))\mathrm{OPT}$, while requiring essentially just $O(md \log^3 m/\alpha)$ runtime.

****

Neural Implicit Shape Editing using Boundary Sensitivity
Arturs Berzins,Moritz Ibing,Leif Kobbelt
Neural fields are receiving increased attention as a geometric representation due to their ability to compactly store detailed and smooth shapes and easily undergo topological changes. Compared to classic geometry representations, however, neural representations do not allow the user to exert intuitive control over the shape. Motivated by this, we leverage boundary sensitivity to express how perturbations in parameters move the shape boundary. This allows us to interpret the effect of each learnable parameter and study achievable deformations. With this, we perform geometric editing: finding a parameter update that best approximates a globally prescribed deformation. Prescribing the deformation only locally allows the rest of the shape to change according to some prior, such as semantics or deformation rigidity. Our method is agnostic to the model and its training and updates the NN in-place. Furthermore, we show how boundary sensitivity helps to optimize and constrain objectives (such as surface area and volume), which are difficult to compute without first converting to another representation, such as a mesh.

****

Amortised Invariance Learning for Contrastive Self-Supervision
Ruchika Chavhan,Jan Stuehmer,Calum Heggan,Mehrdad Yaghoobi,Timothy Hospedales

Contrastive self-supervised learning methods famously produce high quality transferable representations by learning invariances to different data augmentations. Invariances established during pre-training can be interpreted as strong inductive biases. However these may or may not be helpful, depending on if they match the invariance requirements of downstream tasks or not. This has led to several attempts to learn task-specific invariances during pre-training, however, these methods are highly compute intensive and tedious to train. We introduce the notion of amortized invariance learning for contrastive self supervision. In the pre-training stage, we parameterize the feature extractor by differentiable invariance hyper-parameters that control the invariances encoded by the representation. Then, for any downstream task, both linear readout and task-specific invariance requirements can be efficiently and effectively learned by gradient-descent. We evaluate the notion of amortized invariances for contrastive learning over two different modalities: vision and audio, on two widely-used contrastive learning methods in vision: SimCLR and MoCo-v2 with popular architectures like ResNets and Vision Transformers, and SimCLR with ResNet-18 for audio. We show that our amortized features provide a reliable way to learn diverse downstream tasks with different invariance requirements, while using a single feature and avoiding task-specific pre-training. This provides an exciting perspective that opens up new horizons in the field of general purpose representation learning.

****

DIFFUSION GENERATIVE MODELS ON SO(3)

Yesukhei Jagvaral,Francois Lanusse,Rachel Mandelbaum

Diffusion-based generative models represent the current state-of-the-art for image generation. However, standard diffusion models are based on Euclidean geometry and do not translate directly to manifold-valued data. In this work, we develop extensions of both score-based generative models (SGMs) and Denoising Diffusion Probabilistic Models (DDPMs) to the Lie group of 3D rotations, SO(3). SO(3) is of particular interest in many disciplines such as robotics, biochemistry and astronomy/planetary science. Contrary to more general Riemannian manifolds, SO(3) admits a tractable solution to heat diffusion, and allows us to implement efficient training of Diffusion models. We apply both SO(3) DDPMs and SGMs to synthetic densities on SO(3) and demonstrate state-of-the-art results.

****

Certifiably Robust Transformers with 1-Lipschitz Self-Attention

Xiaojun Xu,Linyi Li,Yu Cheng,Subhabrata Mukherjee,Ahmed Hassan Awadallah,Bo Li

Recent works have shown that neural networks with Lipschitz constraints will lead to high adversarial robustness. In this work, we propose the first One-Lipschitz Self-Attention (OLSA) mechanism for Transformer models. In particular, we first orthogonalize all the linear operations in the self-attention mechanism. We then bound the overall Lipschitz constant by aggregating the Lipschitz of each element in the softmax with weighted sum. Based on the proposed self-attention mechanism, we construct an OLSA Transformer to achieve model deterministic certified robustness. We evaluate our model on multiple natural language processing (NLP) tasks and show that it outperforms existing certification on Transformers, especially for models with multiple layers. As an example, for 3-layer Transformers we achieve an ■2 deterministic certified robustness radius of 1.733 and 0.979 on the word embedding space for the Yelp and SST dataset, while the existing SOTA certification baseline of the same embedding space can only achieve 0.061 and 0.110. In addition, our certification is significantly more efficient than previous works, since we only need the output logits and Lipschitz constant for certification. We also fine-tune our OLSA Transformer as a downstream classifier of a pre-trained BERT model and show that it achieves significantly higher certified robustness on BERT embedding space compared with previous works (e.g. from 0.071 to 0.368 on the QQP datasets).

****

Revisiting Populations in multi-agent Communication

Paul Michel,Mathieu Rita,Kory Wallace Mathewson,Olivier Tieleman,Angeliki Lazaridou

Despite evidence from cognitive sciences that larger groups of speakers tend to develop more structured languages in human communication, scaling up to populations has failed to yield significant benefits in emergent multi-agent communication. In this paper we advocate for an alternate population-level training paradigm for referential games based on the idea of "partitioning" the agents into sender-receiver pairs and limiting co-adaptation across pairs. We show that this results in optimizing a different objective at the population level, where agents maximize (1) their respective "internal" communication accuracy and (2) some measure of alignment between agents. In experiments, we find that this leads to the emergence of languages that are significantly more compositional. Moreover, when agents are trained in populations that are not fully connected (ie. not all agent pairs interact at training time), this approach reduces multi-linguality and improves zero-shot communication with new agents (ie. agents are able to communicate successfully with other agents outside their training partners).

****

Univariate vs Multivariate Time Series Forecasting with Transformers
William Michael John Murphy,Ke Chen

Multivariate time series forecasting is a challenging problem and a number of Transformer-based long-term time series forecasting models have been developed to tackle it. These models, however, are impeded by the additional information available in multivariate forecasting. In this paper we propose a simple univariate setting as an alternative method for producing multivariate forecasts. The univariate model is trained on each individual dimension of the time series. This single model is then used to forecast each dimension of the multivariate forecast in turn. A comparative study shows that our setting outperforms state-of-the-art Transformers in the multivariate setting in benchmark datasets. To investigate why, we set three hypotheses and verify them via an empirical study, which leads to a criterion for when our univariate setting is likely to lead to better performance and reveals flaws in the current multivariate Transformers for long-term time series forecasting.

****

Sequential Gradient Coding For Straggler Mitigation
Nikhil Krishnan Muralee Krishnan,MohammadReza Ebrahimi,Ashish J Khisti

In distributed computing, slower nodes (stragglers) usually become a bottleneck. Gradient Coding (GC), introduced by Tandon et al., is an efficient technique that uses principles of error-correcting codes to distribute gradient computation in the presence of stragglers. In this paper, we consider the distributed computation of a sequence of gradients $\{g(1),g(2),\ldots,g(J)\}$, where processing of each gradient $g(t)$ starts in round-t and finishes by round-$(t+T)$. Here $T\geq 0$ denotes a delay parameter. For the GC scheme, coding is only across computing nodes and this results in a solution where $T=0$. On the other hand, having $T>0$ allows for designing schemes which exploit the temporal dimension as well. In this work, we propose two schemes that demonstrate improved performance compared to GC. Our first scheme combines GC with selective repetition of previously unfinished tasks and achieves improved straggler mitigation. In our second scheme, which constitutes our main contribution, we apply GC to a subset of the tasks and repetition for the remainder of the tasks. We then multiplex these two classes of tasks across workers and rounds in an adaptive manner, based on past straggler patterns. Using theoretical analysis, we demonstrate that our second scheme achieves significant reduction in the computational load. In our experiments, we study a practical setting of concurrently training multiple neural networks over an AWS Lambda cluster involving 256 worker nodes, where our framework naturally applies. We demonstrate that the latter scheme can yield a 16\% improvement in runtime over the baseline GC scheme, in the presence of naturally occurring, non-simulated stragglers.

****

TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation
Hyesu Lim,Byeonggeun Kim,Jaegul Choo,Sungha Choi

This paper proposes a novel batch normalization strategy for test-time adaptation. Recent test-time adaptation methods heavily rely on the modified batch normal ization, i.e., transductive batch normalization (TBN), which calculates the mean and the variance from the current test batch rather than using the running mean and variance obtained from the source data, i.e., conventional batch normalization (CBN). Adopting TBN that employs test batch statistics mitigates the performance degradation caused by the domain shift. However, re-estimating normalization statistics using test data depends on impractical assumptions that a test batch should be large enough and be drawn from i.i.d. stream, and we observed that the previous methods with TBN show critical performance drop without the assumptions. In this paper, we identify that CBN and TBN are in a trade-off relationship and present a new test-time normalization (TTN) method that interpolates the statistics by adjusting the importance between CBN and TBN according to the domain-shift sensitivity of each BN layer. Our proposed TTN improves model robustness to shifted domains across a wide range of batch sizes and in various realistic evaluation scenarios. TTN is widely applicable to other test-time adaptation methods that rely on updating model parameters via backpropagation. We demonstrate that adopting TTN further improves their performance and achieves state-of-the-art performance in various standard benchmarks.

****

Choreographer: Learning and Adapting Skills in Imagination

Pietro Mazzaglia,Tim Verbelen,Bart Dhoedt,Alexandre Lacoste,Sai Rajeswar

Unsupervised skill learning aims to learn a rich repertoire of behaviors without external supervision, providing artificial agents with the ability to control and influence the environment. However, without appropriate knowledge and exploration, skills may provide control only over a restricted area of the environment, limiting their applicability. Furthermore, it is unclear how to leverage the learned skill behaviors for adapting to downstream tasks in a data-efficient manner. We present Choreographer, a model-based agent that exploits its world model to learn and adapt skills in imagination. Our method decouples the exploration and skill learning processes, being able to discover skills in the latent state space of the model. During adaptation, the agent uses a meta-controller to evaluate and adapt the learned skills efficiently by deploying them in parallel in imagination. Choreographer is able to learn skills both from offline data, and by collecting data simultaneously with an exploration policy. The skills can be used to effectively adapt to downstream tasks, as we show in the URL benchmark, where we outperform previous approaches from both pixels and states inputs. The skills also explore the environment thoroughly, finding sparse rewards more frequently, as shown in goal-reaching tasks from the DMC Suite and Meta-World.

Project website: https://skillchoreographer.github.io/

****

Disentanglement of Correlated Factors via Hausdorff Factorized Support

Karsten Roth,Mark Ibrahim,Zeynep Akata,Pascal Vincent,Diane Bouchacourt

A grand goal in deep learning research is to learn representations capable of generalizing across distribution shifts.

Disentanglement is one promising direction aimed at aligning a model's representation with the underlying factors generating the data (e.g. color or background). Existing disentanglement methods, however, rely on an often unrealistic assumption: that factors are statistically independent. In reality, factors (like object color and shape) are correlated. To address this limitation, we consider the use of a relaxed disentanglement criterion -- the Hausdorff Factorized Support (HFS) criterion -- that encourages only pairwise factorized support, rather than a factorial distribution, by minimizing a Hausdorff distance. This allows for arbitrary distributions of the factors over their support, including correlations between them. We show that the use of HFS consistently facilitates disentanglement and recovery of ground-truth factors across a variety of correlation settings and benchmarks, even under severe training correlations and correlation shifts, with in parts over +60% in relative improvement over existing disentanglement methods. In addition, we find that leveraging HFS for representation learning can even facilitate transfer to downstream tasks such as classification under distribution shifts. We hope our original approach and positive empirical results inspire further progress on the open problem of robust generalization. Code availab le at https://github.com/facebookresearch/disentangling-correlated-factors.

****

On the optimization and generalization of overparameterized implicit neural networks

Tianxiang Gao,Hongyang Gao

Implicit neural networks have become increasingly attractive in the machine learning community since they can achieve competitive performance but use much less computational resources. Recently, a line of theoretical works established the global convergences for first-order methods such as gradient descent if the implicit networks are over-parameterized. However, as they train all layers together, their analyses are equivalent to only studying the evolution of the output layer. It is unclear how the implicit layer contributes to the training. Thus, in this paper, we restrict ourselves to only training the implicit layer. We show that global convergence is guaranteed, even if only the implicit layer is trained. On the other hand, the theoretical understanding of when and how the training performance of an implicit neural network can be generalized to unseen data is still under-explored. Although this problem has been studied in standard feed-forward networks, the case of implicit neural networks is still intriguing since implicit networks theoretically have infinitely many layers. Therefore, this paper investigates the generalization error for implicit neural networks. Specifically, we study the generalization of an implicit network activated by the ReLU function over random initialization. We provide a generalization bound that is initialization sensitive. As a result, we show that gradient flow with proper random initialization can train a sufficient over-parameterized implicit network to achieve arbitrarily small generalization errors.

****

Differentially Private Conditional Text Generation For Synthetic Data Production

Pranav Putta,Ander Steele,Joseph W Ferrara

Companies have faced increasing pressure in recent years to anonymize user collected data when sharing internally or to third parties. Text data in particular contains copious amounts of personally identifiable information that has proven to be difficult to de-identify while remain useful for the party of interest. Previous works have suggested that synthetic text generation could provide a promising avenue to curate high performant and private datasets. In this paper, we introduce an approach to synthesize high utility text classification datasets by performing conditional generation through a large language model, distilGPT2, while providing measurable guarantees via differential privacy. We show that naive approaches suffer heavily from utility loss by entangling task-relevant factors in the transformer embedding space, making controlled generation more difficult. We analyze how incorporating a secondary learning objective can improve the performance of the generative model, improving utility of the generated data.

****

Multi-Task Structural Learning using Local Task Similarity induced Neuron Creation and Removal

Naresh Kumar Gurulingan,Elahe Arani,Bahram Zonooz

Multi-task learning has the potential to improve generalization by maximizing positive transfer between tasks while reducing task interference. Fully achieving this potential is hindered by manually designed architectures that remain static throughout training. In contrast, learning in the brain occurs through structural changes that are in tandem with changes in synaptic strength. Therefore, we propose Multi-Task Structural Learning (MTSL) which simultaneously learns the multi-task architecture and its parameters. MTSL begins with an identical single task network for each task and alternates between a task learning phase and a structural learning phase. In the task learning phase, each network specializes in the corresponding task. In each of the structural learning phases, starting from the earliest layer, locally similar task layers first transfer their knowledge to a newly created group layer after which they become redundant and are removed. Our experimental results show that MTSL achieves competitive generalization with various baselines and improves robustness to out-of-distribution data.

****

Generating Sequences by Learning to Self-Correct

Sean Welleck,Ximing Lu,Peter West,Faeze Brahman,Tianxiao Shen,Daniel Khashabi,Yejin Choi

Sequence generation applications require satisfying semantic constraints, such as ensuring that programs are correct, using certain keywords, or avoiding undesirable content. Language models, whether fine-tuned or prompted with few-shot demonstrations, frequently violate these constraints, and lack a mechanism to iteratively revise their outputs. Moreover, some powerful language models are of extreme scale or inaccessible, making it inefficient, if not infeasible, to update their parameters for task-specific adaptation. We present Self-Correction, an approach that decouples an imperfect base generator (an off-the-shelf language model or supervised sequence-to-sequence model) from a separate corrector that learns to iteratively correct imperfect generations. To train the corrector, we propose an online training procedure that can use either scalar or natural language feedback on intermediate imperfect generations. We show that Self-Correction improves upon the base generator in three diverse generation tasks - mathematical program synthesis, lexically-constrained generation, and toxicity control - even when the corrector is much smaller than the base generator.

****

Bringing robotics taxonomies to continuous domains via GPLVM on hyperbolic manifolds

Noémie Jaquier,Leonel Rozo,Miguel González-Duque,Viacheslav Borovitskiy,Tamim Asfour

Robotic taxonomies have appeared as high-level hierarchical abstractions that classify how humans move and interact with their environment. They have proven useful to analyse grasps, manipulation skills, and whole-body support poses. Despite the efforts devoted to design their hierarchy and underlying categories, their use in application fields remains scarce. This may be attributed to the lack of computational models that fill the gap between the discrete hierarchical structure of the taxonomy and the high-dimensional heterogeneous data associated to its categories. To overcome this problem, we propose to model taxonomy data via hyperbolic embeddings that capture the associated hierarchical structure. To do so, we formulate a Gaussian process hyperbolic latent variable model and enforce the taxonomy structure through graph-based priors on the latent space and distance-preserving back constraints. We test our model on the whole-body support pose taxonomy to learn hyperbolic embeddings that comply with the original graph structure. We show that our model properly encodes unseen poses from existing or new taxonomy categories, it can be used to generate trajectories between the embeddings, and it outperforms its Euclidean counterparts.

****

COC curve: operating neural networks at high accuracy and low manual effort

Sara Sangalli,Ertunc Erdil,Ender Konukoglu

In human-AI collaboration systems for critical applications based on neural networks, humans should set an operating point based on a model's confidence to determine when the decision should be delegated to experts.

The underlying assumption is that the network's confident predictions are also correct.

However, modern neural networks are notoriously overconfident in their predictions, thus they achieve lower accuracy even when operated at high confidence. Network calibration methods mitigate this problem by encouraging models to make predictions whose confidence is consistent with the accuracy, i.e., encourage confidence to reflect the number of mistakes the network is expected to make. However, they do not consider that data need to be manually analysed by experts in critical applications if the confidence of the network is below a certain level. This can be crucial for applications where available expert time is limited and expensive, e.g., medical ones.

In this paper, we propose (1) Confidence Operating Characteristics (COC) curve that assesses a predictive model in terms of accuracy and manual analysis it requires for varying operating points on confidence, and (2) a new loss function for classification that takes into account both aspects and derived from the COC curve.

We perform extensive experiments on multiple computer vision and medical image datasets for classification and compare the proposed approach with the existing network calibration methods. Our results demonstrate that our method improves classification accuracy while delegating less number of decisions to human experts, achieves better out-of-distribution samples detection and on par calibration performance compared to existing methods.

****

Repository-Level Prompt Generation for Large Language Models of Code

Disha Shrivastava,Hugo Larochelle,Daniel Tarlow

With the success of large language models (LLMs) of code and their use as code assistants (e.g.\ Codex used in GitHub Copilot, techniques for introducing domain-specific knowledge in the prompt design process become important. In this work, we propose a framework called Repo-Level Prompt Generator that learns to generate example-specific prompts using prompt proposals. The prompt proposals take context from the entire repository, thereby incorporating both the structure of the repository and the context from other relevant files (e.g.\ imports, parent class files). Our technique doesn't require any access to the weights of the LLM, making it applicable in cases where we only have black-box access to the LLM. We conduct experiments on the task of single-line code-autocompletion using code repositories taken from Google Code archives. We demonstrate that an oracle constructed from our prompt proposals gives a remarkably high relative improvement of 36\% over Codex, showing the quality of these proposals. Further, we show that when we train a model to select the best prompt proposal, we can achieve significant performance gains over Codex and other baselines.

****

FP_AINet: Fusion Prototype with Adaptive Induction Network for Few-Shot Learning

Mengping Dong,xue Liu,Chaojun Cen,YANG MI,Zhenbo Li

Conventional prototypical network treats all samples equally and does not consider the effects of noisy samples, which leads to a biased class representation. In this paper, we propose a novel Fusion Prototype with Adaptive Induction Network (FP_AINet) for few-shot learning that can learn representative prototypes from a few support samples. Specifically, to address the problem of noisy samples in the support set, an adaptive induction network is developed, which can learn different class representations for diverse queries and assign adaptive scores for support samples according to their relative significance. Moreover, the proposed model can generate a more accurate prototype than comparison methods by considering the query-related samples. With an increasing of samples, the prototypical network is more expressive since the Adaptive Induction Network ignores the relative local features. As a result, a Gaussian-based fusion algorithm is designed to learn more representative prototypes. Extensive experiments are conducted on three datasets: miniImageNet, tieredImageNet, and CIFAR_FS. The experimental results compared with the state-of-the-art few-shot learning methods demonstrate the superiority of FP_AINet.

****

Neural Network Differential Equation Solvers allow unsupervised error estimation and correction

Akshunna S. Dogra,Jeffrey B. Lai,Min Chul Lee,Martin Peev

Neural Network Differential Equation (NN DE) solvers have surged in popularity due to a combination of factors: computational advances making their optimization more tractable, their capacity to handle high dimensional problems, easy interpretability, etc. However, most NN DE solvers suffer from a fundamental limitation: their loss functions are not explicitly dependent on the errors associated with the solution estimates. As such, validation and error estimation usually requires knowledge of the true solution. Indeed, when the true solution is unknown, we are often reduced to simply hoping that a ``\textit{low enough}'' loss implies ``\textit{small enough}'' errors, since explicit relationships between the two are not available. In this work, we describe a general strategy for efficiently constructing error estimates and corrections for Neural Network Differential Equation solvers. Our methods do not require \textit{a priori} knowledge of the true solutions and obtain explicit relationships between loss functions and the errors, given certain assumptions on the DE. In turn, these explicit relationships directly allow us to estimate and correct for the errors.

****

Wide Attention is the Way Forward for Transformers

Jason Ross Brown,Yiren Zhao,Ilia Shumailov,Robert D. Mullins

The Transformer is an extremely powerful and prominent deep learning architecture. In this work, we challenge the commonly held belief in deep learning that going deeper is better, and show an alternative design approach that is building wider attention Transformers. We demonstrate that wide single layer Transformer models can compete with or outperform deeper ones in a variety of Natural Language Processing (NLP) tasks when both are trained from scratch. The impact of changing the model aspect ratio on Transformers is then studied systematically. This ratio balances the number of layers and the number of attention heads per layer while keeping the total number of attention heads and all other hyperparameters constant. On average, across 4 NLP tasks and 10 attention types, single layer wide models perform 0.3% better than their deep counterparts. We show an in-depth evaluation and demonstrate how wide models require a far smaller memory footprint and can run faster on commodity hardware, in addition, these wider models are also more interpretable. For example, a single layer Transformer on the IMDb byte level text classification has 3.1x faster inference latency on a CPU than its equally accurate deeper counterpart, and is half the size. Our results suggest that the critical direction for building better Transformers for NLP is their width, and that their depth is less relevant.

****

Variational Prompt Tuning Improves Generalization of Vision-Language Models

Mohammad Mahdi Derakhshani,Enrique Sanchez,Adrian Bulat,Victor Guilherme Turrisi da Costa,Cees G. M. Snoek,Georgios Tzimiropoulos,Brais Martinez

Prompt tuning provides an efficient mechanism to adapt large vision-language models to downstream tasks by treating part of the input language prompts as learnable parameters while freezing the rest of the model. Existing works for prompt tuning are however prone to damaging the generalization capabilities of the foundation models, because the learned prompts lack the capacity of covering certain concepts within the language model. To avoid such limitation, we propose a probabilistic modeling of the underlying distribution of prompts, allowing prompts within the support of an associated concept to be derived through stochastic sampling. This results in a more complete and richer transfer of the information captured by the language model, providing better generalization capabilities for downstream tasks. The resulting algorithm relies on a simple yet powerful variational framework that can be directly integrated with other developments. We show our approach is seamlessly integrated into both standard and conditional prompt learning frameworks, improving the performance on both cases considerably, especially with regards to preserving the generalization capability of the original model. Our method provides the current state-of-the-art for prompt learning, surpassing CoCoOp by 1.6% average Top-1 accuracy on the standard benchmark. Remarkably, it even surpasses the original CLIP model in terms of generalization to new classes. Implementation code will be released.

****

DCT-DiffStride: Differentiable Strides with Real-Valued Data

Clayton Harper,Mitchell Thornton,Eric Larson

Reducing the size of intermediate feature maps within various neural network architectures is critical for generalization performance, and memory and computational complexity. Until recently, most methods required downsampling rates (i.e., decimation) to be predefined and static during training, with optimal downsampling rates requiring a vast hyper-parameter search. Recent work has proposed a novel and differentiable method for learning strides named DiffStride which uses the discrete Fourier transform (DFT) to learn strides for decimation. However, in many cases the DFT does not capture signal properties as efficiently as the disc rete cosine transform (DCT). Therefore, we propose an alternative method for learning decimation strides, DCT-DiffStride, as well as new regularization methods to reduce model complexity. Our work employs the DCT and its inverse as a low-pass filter in the frequency domain to reduce feature map dimensionality. Leveraging the well-known energy compaction properties of the DCT for natural signals, we evaluate DCT-DiffStride with its competitors on image and audio datasets demonstrating a favorable tradeoff in model performance and model complexity compared to competing methods. Additionally, we show DCT-DiffStride and DiffStride can be applied to data outside the natural signal domain, increasing the general applications of such methods.

****

Interneurons accelerate learning dynamics in recurrent neural networks for statistical adaptation

David Lipshutz,Cengiz Pehlevan,Dmitri Chklovskii

Early sensory systems in the brain rapidly adapt to fluctuating input statistics, which requires recurrent communication between neurons. Mechanistically, such recurrent communication is often indirect and mediated by local interneurons. In this work, we explore the computational benefits of mediating recurrent communication via interneurons compared with direct recurrent connections. To this end, we consider two mathematically tractable recurrent neural networks that statistically whiten their inputs --- one with direct recurrent connections and the other with interneurons that mediate recurrent communication. By analyzing the corresponding continuous synaptic dynamics and numerically simulating the networks, we show that the network with interneurons is more robust to initialization than the network with direct recurrent connections in the sense that the convergence time for the synaptic dynamics in the network with interneurons (resp. direct recurrent connections) scales logarithmically (resp. linearly) with the spectrum of their initialization. Our results suggest that interneurons are computationally useful for rapid adaptation to changing input statistics. Interestingly, the network with interneurons is an overparameterized solution of the whitening objective for the network with direct recurrent connections, so our results can be viewed as a recurrent neural network analogue of the implicit acceleration phenomenon observed in overparameterized feedforward linear networks.

****

Understanding DDPM Latent Codes Through Optimal Transport

Valentin Khrulkov,Gleb Ryzhakov,Andrei Chertkov,Ivan Oseledets

Diffusion models have recently outperformed alternative approaches to model the distribution of natural images. Such diffusion models allow for deterministic sampling via the probability flow ODE, giving rise to a latent space and an encoder map. While having important practical applications, such as the estimation of the likelihood, the theoretical properties of this map are not yet fully understood. In the present work, we partially address this question for the popular case of the VP-SDE (DDPM) approach. We show that, perhaps surprisingly, the DDPM encoder map coincides with the optimal transport map for common distributions; we support this claim by extensive numerical experiments using advanced tensor train solver for multidimensional Fokker-Planck equation. We provide additional theoretical evidence for the case of multivariate normal distributions.

****

Soft Sampling for Efficient Training of Deep Neural Networks on Massive Data

Xiaodong Cui,Ashish Mittal,Songtao Lu,Wei Zhang,George Saon,Brian Kingsbury

We investigate soft sampling which is a simple yet effective approach for efficient training of large-scale deep neural network models when dealing with massive data. Soft sampling selects a subset uniformly at random with replacement from the full data set in each epoch. First, we derive a theoretical convergence guarantee for soft sampling on non-convex objective functions and give the convergence rate. Next, we analyze the data coverage and occupancy properties of soft sampling from the perspective of the coupon collector's problem. And finally, we evaluate soft sampling on various machine learning tasks using various network architectures and demonstrate its effectiveness. Compared to existing coreset-based data selection methods, soft sampling offers a better accuracy-efficiency trade

-off. Especially on real-world industrial scale data sets, soft sampling can achieve significant speedup and competitive performance with almost no additional computing cost.

****

Learning About Progress From Experts

Jake Bruce,Ankit Anand,Bogdan Mazoure,Rob Fergus

Many important tasks involve some notion of long-term progress in multiple phases: e.g. to clean a shelf it must be cleared of items, cleaning products applied, and then the items placed back on the shelf. In this work, we explore the use of expert demonstrations in long-horizon tasks to learn a monotonically increasing function that summarizes progress. This function can then be used to aid agent exploration in environments with sparse rewards. As a case study we consider the NetHack environment, which requires long-term progress at a variety of scales and is far from being solved by existing approaches. In this environment, we demonstrate that by learning a model of long-term progress from expert data containing only observations, we can achieve efficient exploration in challenging sparse tasks, well beyond what is possible with current state-of-the-art approaches. We have made the curated gameplay dataset used in this work available at https://github.com/deepmind/nao_top10.

****

Learning Fair Graph Representations via Automated Data Augmentations

Hongyi Ling,Zhimeng Jiang,Youzhi Luo,Shuiwang Ji,Na Zou

We consider fair graph representation learning via data augmentations. While this direction has been explored previously, existing methods invariably rely on certain assumptions on the properties of fair graph data in order to design fixed strategies on data augmentations. Nevertheless, the exact properties of fair graph data may vary significantly in different scenarios. Hence, heuristically designed augmentations may not always generate fair graph data in different application scenarios. In this work, we propose a method, known as Graphair, to learn fair representations based on automated graph data augmentations. Such fairness-aware augmentations are themselves learned from data. Our Graphair is designed to automatically discover fairness-aware augmentations from input graphs in order to circumvent sensitive information while preserving other useful information. Experimental results demonstrate that our Graphair consistently outperforms many baselines on multiple node classification datasets in terms of fairness-accuracy trade-off performance. In addition, results indicate that Graphair can automatically learn to generate fair graph data without prior knowledge on fairness-relevant graph properties.

****

A new photoreceptor-inspired CNN layer enables deep learning models of retina to generalize across lighting conditions

Saad Idrees,Greg D Field,Frederick Rieke,Joel Zylberberg

As we move our eyes, and as lighting changes in our environment, the light intensity reaching our retinas changes dramatically and on multiple timescales. Despite these changing conditions, our retinas effortlessly extract visual information that allows downstream brain areas to make sense of the visual world. Such processing capabilities are desirable in many settings, including computer vision systems that operate in dynamic lighting environments like in self-driving cars, and in algorithms that translate visual inputs into neural signals for use in vision-restoring prosthetics. To mimic retinal processing, we first require models that can predict retinal ganglion cell (RGC) responses reliably. While existing state-of-the-art deep learning models can accurately predict RGC responses to visual scenes under steady-state lighting conditions, these models fail under dynamic lighting conditions. This is because changes in lighting markedly alter RGC responses: adaptation mechanisms dynamically tune RGC receptive fields on multiple timescales. Because current deep learning models of the retina have no in-built notion of light level or these adaptive mechanisms, they are unable to accurately predict RGC responses under lighting conditions that they were not trained on. We present here a new deep learning model of the retina that can predict RGC responses to visual scenes at different light levels without requiring trainin g at each light level. Our model combines a fully trainable biophysical front end capturing the fast and slow adaptation mechanisms in the photoreceptors with convolutional neural networks (CNNs) capturing downstream retinal processing. We tested our model’s generalization performance across light levels using monkey and rat retinal data. Whereas conventional CNN models without the photoreceptor layer failed to predict RGC responses when the lighting conditions changed, our model with the photoreceptor layer as a front end fared much better in this challenge. Overall, our work demonstrates a new hybrid approach that equips deep learning models with biological vision mechanisms enabling them to adapt to dynamic environments.

****

3D Neural Embedding Likelihood for Robust Sim-to-Real Transfer in Inverse Graphics

Guangyao Zhou,Nishad Gothoskar,Lirui Wang,Joshua B. Tenenbaum,Dan Gutfreund,Miguel Lazaro-Gredilla,Dileep George,Vikash Mansinghka

A central challenge in 3D scene perception via inverse graphics is robustly modelling the gap between 3D graphics and real-world data. We propose a novel 3D Neural Embedding Likelihood (3DNEL) over RGB-D images to address this gap. 3DNEL uses neural embeddings to predict 2D-3D correspondences from RGB and combines this with depth in a principled manner. 3DNEL is trained entirely from synthetic images and generalizes to real-world data. To showcase this capability, we develop a multi-stage inverse graphics pipeline that uses 3DNEL for 6D object pose estimation from real RGB-D images. Our method outperforms the previous state-of-the-art in sim-to-real pose estimation on the YCB-Video dataset, and improves robustness, with significantly fewer large-error predictions. Unlike existing bottom-up, discriminative approaches that are specialized for pose estimation, 3DNEL adopts a probabilistic generative formulation that jointly models multi-object scenes. This generative formulation enables easy extension of 3DNEL to additional tasks like object and camera tracking from video, using principled inference in the same probabilistic model without task specific retraining.

****

Dynamic Scheduled Sampling with Imitation Loss for Neural Text Generation

Xiang Lin,Prathyusha Jwalapuram,Shafiq Joty

State-of-the-art neural text generation models are typically trained to maximize the likelihood of each token in the ground-truth sequence conditioned on the previous target tokens. However, during inference, the model needs to make a prediction conditioned on the tokens generated by itself. This train-test discrepancy is referred to as exposure bias. Scheduled sampling is a curriculum learning strategy that gradually exposes the model to its own predictions during training to mitigate this bias. Most of the proposed approaches design a scheduler based on training steps, which generally requires careful tuning depending on the training setup. In this work, we introduce Dynamic Scheduled Sampling with Imitation Loss (DySI), which maintains the schedule based solely on the training time accuracy, while enhancing the curriculum learning by introducing an imitation loss, which attempts to make the behavior of the decoder indistinguishable from the behavior of a teacher-forced decoder. DySI is universally applicable across training setups with minimal tuning. Extensive experiments and analysis show that DySI not only achieves notable improvements on standard machine translation benchmarks, but also significantly improves the robustness of other text generation models.

****

Emergence of Maps in the Memories of Blind Navigation Agents

Erik Wijmans,Manolis Savva,Irfan Essa,Stefan Lee,Ari S. Morcos,Dhruv Batra

Animal navigation research posits that organisms build and maintain internal spa- tial representations, or maps, of their environment. We ask if machines – specifically, artificial intelligence (AI) navigation agents – also build implicit (or ‘mental’) maps. A positive answer to this question would (a) explain the surprising phenomenon in recent literature of ostensibly map-free neural-networks achieving strong performance, and (b) strengthen the evidence of mapping as a fundamental mechanism for navigation by intelligent embodied agents, whether they be

biological or artificial. Unlike animal navigation, we can judiciously design the agent’s perceptual system and control the learning paradigm to nullify alternative navigation mechanisms. Specifically, we train ‘blind’ agents – with sensing limited to only egomotion and no other sensing of any kind – to perform PointGoal navigation (‘go to Δx, Δy’) via reinforcement learning. Our agents are composed of navigation-agnostic components (fully-connected and recurrent neural networks), and our experimental setup provides no inductive bias towards mapping. Despite these harsh conditions, we find that blind agents are (1) surprisingly effective navigators in new environments (~95% success); (2) they utilize memory over long horizons (remembering ~1,000 steps of past experience in an episode); (3) this memory enables them to exhibit intelligent behavior (following walls, detecting collisions, taking shortcuts); (4) there is emergence of maps and collision detection neurons in the representations of the environment built by a blind agent as it navigates; and (5) the emergent maps are selective and task dependent (e.g. the agent ‘forgets’ exploratory detours). Overall, this paper presents no new techniques for the AI audience, but a surprising finding, an insight, and an explanation.
****
Latent Neural ODEs with Sparse Bayesian Multiple Shooting
Valerii Iakovlev,Cagatay Yildiz,Markus Heinonen,Harri Lähdesmäki
Training dynamic models, such as neural ODEs, on long trajectories is a hard problem that requires using various tricks, such as trajectory splitting, to make model training work in practice. These methods are often heuristics with poor theoretical justifications, and require iterative manual tuning. We propose a principled multiple shooting technique for neural ODEs that splits the trajectories into manageable short segments, which are optimized in parallel, while ensuring probabilistic control on continuity over consecutive segments. We derive variational inference for our shooting-based latent neural ODE models and propose amortized encodings of irregularly sampled trajectories with a transformer-based recognition network with temporal attention and relative positional encoding. We demonstrate efficient and stable training, and state-of-the-art performance on multiple large-scale benchmark datasets.
****
$\mathcal{O}$-GNN: incorporating ring priors into molecular modeling
Jinhua Zhu,Kehan Wu,Bohan Wang,Yingce Xia,Shufang Xie,Qi Meng,Lijun Wu,Tao Qin,Wengang Zhou,Houqiang Li,Tie-Yan Liu
Cyclic compounds that contain at least one ring play an important role in drug design. Despite the recent success of molecular modeling with graph neural networks (GNNs), few models explicitly take rings in compounds into consideration, consequently limiting the expressiveness of the models. In this work, we design a new variant of GNN, ring-enhanced GNN ($\mathcal{O}$-GNN), that explicitly models rings in addition to atoms and bonds in compounds. In $\mathcal{O}$-GNN, each ring is represented by a latent vector, which contributes to and is iteratively updated by atom and bond representations. Theoretical analysis shows that $\mathcal{O}$-GNN is able to distinguish two isomorphic subgraphs lying on different rings using only one layer while conventional graph convolutional neural networks require multiple layers to distinguish, demonstrating that $\mathcal{O}$-GNN is more expressive. Through experiments, $\mathcal{O}$-GNN shows good performance on $\bf{11}$ public datasets. In particular, it achieves state-of-the-art validation result on the PCQM4Mv1 benchmark (outperforming the previous KDDCup champion solution) and the drug-drug interaction prediction task on DrugBank. Furthermore, $\mathcal{O}$-GNN outperforms strong baselines (without modeling rings) on the molecular property prediction and retrosynthesis prediction tasks.
****
MACTA: A Multi-agent Reinforcement Learning Approach for Cache Timing Attacks and Detection
Jiaxun Cui,Xiaomeng Yang,Mulong Luo,Geunbae Lee,Peter Stone,Hsien-Hsin S. Lee,Benjamin Lee,G. Edward Suh,Wenjie Xiong,Yuandong Tian
Security vulnerabilities in computer systems raise serious concerns as computers process an unprecedented amount of private and sensitive data today. Cache timi

ng attacks (CTA) pose an important practical threat as they can effectively breach many protection mechanisms in today’s systems. However, the current detection techniques for cache timing attacks heavily rely on heuristics and expert knowledge, which can lead to brittleness and the inability to adapt to new attacks. To mitigate the CTA threat, we propose MACTA, a multi-agent reinforcement learning (MARL) approach that leverages population-based training to train both attackers and detectors. Following best practices, we develop a realistic simulated MARL environment, MA-AUTOCAT, which enables training and evaluation of cache-timing attackers and detectors. Our empirical results suggest that MACTA is an effective solution without any manual input from security experts. MACTA detectors can generalize to a heuristic attack not exposed in training with a 97.8% detection rate and reduce the attack bandwidth of adaptive attackers by 20% on average. In the meantime, MACTA attackers are qualitatively more effective than other attacks studied, and the average evasion rate of MACTA attackers against an unseen state-of-the-art detector can reach up to 99%. Furthermore, we found that agents equipped with a Transformer encoder can learn effective policies in situations when agents with multi-layer perceptron encoders do not in this environment, suggesting the potential of Transformer structures in CTA problems.

****

Training Normalizing Flows from Dependent Data

Matthias Kirchler,Christoph Lippert,Marius Kloft

Normalizing flows are powerful non-parametric statistical models that function as a hybrid between density estimators and generative models. Current learning algorithms for normalizing flows assume that data points are sampled independently, an assumption that is frequently violated in practice, which may lead to erroneous density estimation and data generation. We propose a likelihood objective of normalizing flows incorporating dependencies between the data points, for which we derive a flexible and efficient learning algorithm suitable for different dependency structures. We show that respecting dependencies between observations can improve empirical results on both synthetic and real-world data.

****

Spectral Augmentation for Self-Supervised Learning on Graphs

Lu Lin,Jinghui Chen,Hongning Wang

Graph contrastive learning (GCL), as an emerging self-supervised learning technique on graphs, aims to learn representations via instance discrimination. Its performance heavily relies on graph augmentation to reflect invariant patterns that are robust to small perturbations; yet it still remains unclear about what graph invariance GCL should capture. Recent studies mainly perform topology augmentations in a uniformly random manner in the spatial domain, ignoring its influence on the intrinsic structural properties embedded in the spectral domain. In this work, we aim to find a principled way for topology augmentations by exploring the invariance of graphs from the spectral perspective. We develop spectral augmentation which guides topology augmentations by maximizing the spectral change. Extensive experiments on both graph and node classification tasks demonstrate the effectiveness of our method in self-supervised representation learning. The proposed method also brings promising generalization capability in transfer learning, and is equipped with intriguing robustness property under adversarial attacks. Our study sheds light on a general principle for graph topology augmentation.

****

An ensemble view on mixup

David Lopez-Paz,Ishmael Belghazi,Diane Bouchacourt,Elvis Dohmatob,Badr Youbi Idrissi,Levent Sagun,Andrew M Saxe

Deep ensembles are widely used to improve the generalization, calibration, uncertainty estimates and adversarial robustness of neural networks. In parallel, the data augmentation technique of mixup has grown popular for the very same reasons. Could these two techniques be related? This work suggests that both implement a similar inductive bias to “linearize” decision boundaries. We show how to obt

ain diverse predictions from a single mixup machine by interpolating a test instance with multiple reference points. These “mixup ensembles” are cheap: one needs to train and store one single model, as opposed to the K independent members forming a deep ensemble. Motivated by the limitations of ensembles to model uncertainty far away from the training data, we propose a variant of mixup that builds augmented examples using both random interpolations and extrapolations of examples. We evaluate the efficacy of our proposed methods across a variety of in-domain and out-domain metrics on the CIFAR-10 and CIFAR-10-NEG datasets.
****

Improving Adversarial Robustness by Contrastive Guided Diffusion Process
Yidong Ouyang,Liyan Xie,Guang Cheng

Synthetic data generation has become an emerging tool to help improve the adversarial robustness in classification tasks since robust learning requires a significantly larger amount of training samples compared with standard classification tasks. Among various deep generative models, the diffusion model has been shown to produce high-quality synthetic images and has achieved good performance in improving the adversarial robustness. However, diffusion-type methods are typically slow in data generation as compared with other generative models. Although different acceleration techniques have been proposed recently, it is also of great importance to study how to improve the sample efficiency of generated data for the downstream task. In this paper, we first analyze the optimality condition of synthetic distribution for achieving non-trivial robust accuracy. We show that enhancing the distinguishability among the generated data is critical for improving adversarial robustness. Thus, we propose the Contrastive-Guided Diffusion Process (Contrastive-DP), which adopts the contrastive loss to guide the diffusion model in data generation. We verify our theoretical results using simulations and demonstrate the good performance of Contrastive-DP on image datasets.
****

σReparam: Stable Transformer Training with Spectral Reparametrization
Shuangfei Zhai,Tatiana Likhomanenko,Etai Littwin,Jason Ramapuram,Dan Busbridge,Yizhe Zhang,Jiatao Gu,Joshua M. Susskind

Training stability is of great importance to Transformers. In this work, we investigate the training dynamics of Transformers by examining the evolution of the attention layers. In particular, we track the "attention entropy" for each attention head during the course of training, which is a proxy of the attention's sharpness. We observe a common, non monotonic evolution of attention entropy across different settings: the attention entropy first quickly decreases in the initial phase of training, followed by quickly increasing, and finally entering a long stable phase. While the exact shape can be affected by hyperparameters such as warmup, initialization, learning rate etc., we found that there is a close correlation between the minima of attention entropy and the model's training stability. To this end, we propose a simple and efficient solution dubbed σReparam, where we reparametrize all linear layers with Spectral Normalization and an additional learned scalar. We provide a lower bound on the attention entropy as a function of the spectral norms of the query and key projections, which suggests that small attention entropy can be obtained with large spectral norms. σReparam decouples the growth rate of a weight matrix's spectral norm from its dimensionality, which we verify empirically. We conduct experiments with σReparam on image classification, image self supervised learning, automatic speech recognition and language modeling tasks. We show that σReparam provides great stability and robustness with respect to the choice of hyperparameters.
****

PAC Reinforcement Learning for Predictive State Representations
Wenhao Zhan,Masatoshi Uehara,Wen Sun,Jason D. Lee

In this paper we study online Reinforcement Learning (RL) in partially observable dynamical systems. We focus on the Predictive State Representations (PSRs) model, which is an expressive model that captures other well-known models such as Partially Observable Markov Decision Processes (POMDP). PSR represents the states using a set of predictions of future observations and is defined entirely using observable quantities. We develop a novel model-based algorithm for PSRs that c an learn a near optimal policy in sample complexity scaling polynomially with respect to all the relevant parameters of the systems. Our algorithm naturally works with function approximation to extend to systems with potentially large state and observation spaces. We show that given a realizable model class, the sample complexity of learning the near optimal policy only scales polynomially with respect to the statistical complexity of the model class, without any explicit polynomial dependence on the size of the state and observation spaces. Notably, our work is the first work that shows polynomial sample complexities to compete with the globally optimal policy in PSRs. Finally, we demonstrate how our general theorem can be directly used to derive sample complexity bounds for special models including m-step weakly revealing and m-step decodable tabular POMDPs, POMDPs with low-rank latent transition, and POMDPs with linear emission and latent transition.

****

Federated Learning on Adaptively Weighted Nodes by Bilevel Optimization

Yankun Huang,Qihang Lin,Nick Street,Stephen Baek

We propose a federated learning method with weighted nodes in which the weights can be modified to optimize the model’s performance on a separate validation set. The problem is formulated as a bilevel optimization problem where the inner problem is a federated learning problem with weighted nodes and the outer problem focuses on optimizing the weights based on the validation performance of the model returned from the inner problem. A communication-efficient federated optimization algorithm is designed to solve this bilevel optimization problem. We analyze the generalization performance of the output model and identify the scenarios when our method is in theory superior to training a model locally and superior to federated learning with static and evenly distributed weights.

****

Removing Structured Noise with Diffusion Models

Tristan Stevens,Ruud Van Sloun,Jean-luc Robert,Faik C Meral,Jason Yu,Junseob Shin

Solving ill-posed inverse problems requires careful formulation of prior beliefs over the signals of interest and an accurate description of their manifestation into noisy measurements. Handcrafted signal priors based on e.g. sparsity are increasingly replaced by data-driven deep generative models, and several groups have recently shown that state-of-the-art score-based diffusion models yield particularly strong performance and flexibility. In this paper, we show that the powerful paradigm of posterior sampling with diffusion models can be extended to include rich, structured, noise models. To that end, we propose a joint conditional reverse diffusion process with learned scores for the noise and signal-generating distribution. We demonstrate strong performance gains across various inverse problems with structured noise, outperforming competitive baselines that use normalizing flows and adversarial networks. This opens up new opportunities and relevant practical applications of diffusion modeling for inverse problems in the context of non-Gaussian measurements.

****

Stein Variational Goal Generation for adaptive Exploration in Multi-Goal Reinforcement Learning

Nicolas Castanet,Olivier Sigaud,sylvain lamprier

Multi-goal Reinforcement Learning has recently attracted a large amount of research interest. By allowing experience to be shared between related training tasks, this setting favors generalization for new tasks at test time, whenever some smoothness exists in the considered representation space of goals. However, in settings with discontinuities in state or goal spaces (e.g. walls in a maze), a majority of goals are difficult to reach, due to the sparsity of rewards in the absence of expert knowledge. This implies hard exploration, for which some curriculum of goals must be discovered, to help agents learn by adapting training tasks to their current capabilities. We propose a novel approach: Stein Variational Goal Generation (SVGG), which builds on recent automatic curriculum learning techniques for goal-conditioned policies. SVGG seeks at preferably sampling new goals in the zone of proximal development of the agent, by leveraging a learned mode l of its abilities and a goal distribution modeled as particles in the exploration space. Our approach relies on Stein Variational Gradient Descent to dynamically attract the goal sampling distribution in areas of appropriate difficulty. We demonstrate the performances of the approach, in terms of success coverage in the goal space, compared to recent state-of-the-art RL methods for hard exploration problems.

****

Fourier PINNs: From Strong Boundary Conditions to Adaptive Fourier Bases

Madison Cooley,Da Long,Robert Kirby,Shandian Zhe

Interest in Physics-Informed Neural Networks (PINNs) is rising as a mesh-free alternative to traditional numerical solvers for partial differential equations (PDEs). While successful, PINNs often struggle to learn high-frequency and multi-scale target solutions—which, according to prior analysis, might arise from competition during optimization between the weakly enforced boundary loss and residual loss terms. By creatively modifying the neural network architecture, some simple boundary conditions (BCs) can be satisfied exactly without jointly optimizing an additional loss term, thus avoiding the aforementioned competition altogether. Motivated by this analysis, we first study a strong BC version of PINNs for Dirichlet BCs and observe a consistent improvement compared to the standard PINNs. We conducted a Fourier analysis and found that strong BC PINNs can better learn the amplitudes of high-frequency components of the target solutions. While BC PINNs provide a promising improvement, constructing these unique architectures is an intricate process made difficult (if not impossible) by certain BCs and domain geometries. Enlightened by our analysis, we propose Fourier PINNs—a simple, general, yet powerful method that augments PINNs with pre-specified, dense Fourier bases. Our proposed architecture likewise better learns high-frequency components but places no restrictions on the particular BCs. We developed an adaptive learning and basis selection algorithm based on alternating NN basis optimization, Fourier and NN basis coefficient estimations, and coefficient truncation. This schema can flexibly identify the significant frequencies while weakening the nominal to better capture the target solution's power spectrum. We show the advantage of our approach in learning high-frequency and multi-scale solutions in a set of systematic experiments.

****

Distributed Graph Neural Network Training with Periodic Stale Representation Synchronization

Zheng Chai,Guangji Bai,Liang Zhao,Yue Cheng

Despite the recent success of Graph Neural Networks (GNNs), it remains challenging to train a GNN on large graphs with over millions of nodes & billions of edges, which are prevalent in many graph-based applications such as social networks, recommender systems, and knowledge graphs. Traditional sampling-based methods accelerate GNN training by dropping edges and nodes, which impairs the graph integrity and model performance. Differently, distributed GNN algorithms accelerate GNN training by utilizing multiple computing devices and can be classified into two types: "partition-based" methods enjoy low communication cost but suffer from information loss due to dropped edges, while "propagation-based" methods avoid information loss but suffer from prohibitive communication overhead caused by neighbor explosion. To jointly address these problems, this paper proposes DIGEST (DIstributed Graph reprEsentation SynchronizaTion), a novel distributed GNN training framework that synergizes the complementary strength of both categories of existing methods. We propose to allow each device utilize the stale representations of its neighbors in other subgraphs during subgraph parallel training. This way, out method preserves global graph information from neighbors to avoid information loss and reduce the communication cost. Therefore, DIGEST is both computation-efficient and communication-efficient as it does not need to frequently (re-)compute and transfer the massive representation data across the devices, due to neighbor explosion. DIGEST provides synchronous and asynchronous training manners for homogeneous and heterogeneous training environment, respectively. We proved that the approximation error induced by the staleness of the representations can be upper-bounded. More importantly, our convergence analysis demonstrates that DIGEST enjoys the state-of-the-art convergence rate. Extensive experimental evaluation on large, real-world graph datasets shows that DIGEST achieves up to 21.82× speedup without compromising the performance compared to state-of-the-art distributed GNN training frameworks

****

SAGE: Semantic-Aware Global Explanations for Named Entity Recognition

Andrea Zugarini,Leonardo Rigutini

In the last decades, deep learning approaches achieved impressive results in many research fields,
such as Computer Vision and Natural Language Processing (NLP).
NLP in particular has greatly benefit from unsupervised methods that allow to learn distributed representation of language.
On the race for better performances Language Models have reached hundred of billions parameters nowadays.
Despite the remarkable results, deep models are still far from being fully exploited in real world applications.
Indeed, these approaches are black-boxes, i.e. they are not interpretable by design nor explainable, which is often crucial to make decisions in business.
Several task-agnostic methods have been proposed in literature to explain models' decisions.
Most techniques rely on the "local" assumption, i.e. explanations are made example-wise.
In this paper instead, we present a post-hoc method to produce highly interpretable global rules to explain NLP classifiers.
Rules are extracted with a data mining approach on a semantically enriched input representation, instead of using words/wordpieces solely.
Semantic information yields more abstract and general rules that are both more explanatory and less complex, while being also better at reflecting the model behaviour.
In the experiments we focus on Named Entity Recognition, an NLP task where explainability is under-investigated.
We explain the predictions of BERT NER classifiers trained on two popular benchmarks, CoNLL03 and Ontonotes, and compare our model against LIME.

****

Decentralized Optimistic Hyperpolicy Mirror Descent: Provably No-Regret Learning in Markov Games

Wenhao Zhan,Jason D. Lee,Zhuoran Yang

We study decentralized policy learning in Markov games where we control a single agent to play with nonstationary and possibly adversarial opponents. Our goal is to develop a no-regret online learning algorithm that (i) takes actions based on the local information observed by the agent and (ii) is able to find the best policy in hindsight. For such a problem, the nonstationary state transitions due to the varying opponent pose a significant challenge. In light of a recent hardness result (Liu et al., 2022), we focus on the setting where the opponent's previous policies are revealed to the agent for decision making. With such an information structure, we propose a new algorithm, Decentralized Optimistic hypeRpolicy mIrror deScent (DORIS), which achieves $\sqrt{K}$-regret in the context of general function approximation, where K is the number of episodes. Moreover, when all the agents adopt DORIS, we prove that their mixture policy constitutes an approximate coarse correlated equilibrium. In particular, DORIS maintains a hyperpolicy which is a distribution over the policy space. The hyperpolicy is updated via mirror descent, where the update direction is obtained by an optimistic variant of least-squares policy evaluation. Furthermore, to illustrate the power of our method, we apply DORIS to constrained and vector-valued MDPs, which can be formulated as zero-sum Markov games with a fictitious opponent.

****

Graph Contrastive Learning with Model Perturbation

Qiaoyu Tan,Sirui Ding,Ninghao Liu,Soo-Hyun Choi,Li Li,Rui Chen,Xia Hu

Graph contrastive learning (GCL) has achieved great success in pre-training graph neural networks (GNN) without ground-truth labels. The performance of GCL main ly rely on designing high quality contrastive views via data augmentation. However, finding desirable augmentations is difficult and requires cumbersome efforts due to the diverse modalities in graph data. In this work, we study model perturbation to perform efficient contrastive learning on graphs without using data augmentation. Instead of searching for the optimal combination among perturbing nodes, edges or attributes, we propose to conduct perturbation on the model architectures (i.e., GNNs). However, it is non-trivial to achieve effective perturbations on GNN models without performance dropping compared with its data augmentation counterparts. This is because data augmentation 1) makes complex perturbation in the graph space, so it is hard to mimic its effect in the model parameter space with a fixed noise distribution, and 2) has different disturbances even on the same nodes between two views owning to the randomness. Motivated by this, we propose a novel model perturbation framework -- \textsc{PerturbGCL} to pre-train GNN encoders. We focus on perturbing two key operations in a GNN, including message propagation and transformation. Specifically, we propose \emph{weightPrune} to create a dynamic perturbed model to contrast with the target one by pruning its transformation weights according to their magnitudes. Contrasting the two models will lead to adaptive mining of the perturbation distribution from the data. Furthermore, we present \emph{randMP} to disturb the steps of message propagation in two contrastive models. By randomly choosing the propagation steps during training, it helps to increase local variances of nodes between the contrastive views. Despite the simplicity, coupling the two strategies together enable us to perform effective contrastive learning on graphs with model perturbation. We conduct extensive experiments on 15 benchmarks. The results demonstrate the superiority of \textsc{PerturbGCL}: it can achieve competitive results against strong baselines across both node-level and graph-level tasks, while requiring shorter computation time. The code is available at \url{https://anonymous.4open.science/r/PerturbGCL-F17D}.

****

Robust Scheduling with GFlowNets

David W Zhang,Corrado Rainone,Markus Peschl,Roberto Bondesan

Finding the best way to schedule operations in a computation graph is a classical NP-hard problem which is central to compiler optimization. However, evaluating the goodness of a schedule on the target hardware can be very time-consuming. Traditional approaches as well as previous machine learning ones typically optimize proxy metrics, which are fast to evaluate but can lead to bad schedules when tested on the target hardware. In this work, we propose a new approach to scheduling by sampling proportionally to the proxy metric using a novel GFlowNet method. We introduce a technique to control the trade-off between diversity and goodness of the proposed schedules at inference time and demonstrate empirically that the pure optimization baselines can lead to subpar performance with respect to our approach when tested on a target model. Furthermore, we show that conditioning the GFlowNet on the computation graph enables generalization to unseen scheduling problems for both synthetic and real-world compiler datasets.

****

Pareto Manifold Learning: Tackling multiple tasks via ensembles of single-task models

Nikolaos Dimitriadis,Pascal Frossard,François Fleuret

In Multi-Task Learning, tasks may compete and limit the performance achieved on each other rather than guiding the optimization trajectory to a common solution, superior to its single-task counterparts. There is often not a single solution that is optimal for all tasks, leading practitioners to balance tradeoffs between tasks' performance, and to resort to optimality in the Pareto sense. Current Multi-Task Learning methodologies either completely neglect this aspect of functional diversity, and produce one solution in the Pareto Front predefined by their optimization schemes, or produce diverse but discrete solutions, each requiring a separate training run. In this paper, we conjecture that there exist Pareto Subspaces, i.e., weight subspaces where multiple optimal functional solutions lie. We propose Pareto Manifold Learning, an ensembling method in weight space that is able to discover such a parameterization and produces a continuous Pareto Fr ont in a single training run, allowing practitioners to modulate the performance on each task during inference on the fly. We validate the proposed method on a diverse set of multi-task learning benchmarks, ranging from image classification to tabular datasets and scene understanding, and show that Pareto Manifold Learning outperforms state-of-the-art algorithms.

****

Autoregressive Conditional Neural Processes

Wessel Bruinsma,Stratis Markou,James Requeima,Andrew Y. K. Foong,Tom Andersson,Anna Vaughan,Anthony Buonomo,Scott Hosking,Richard E Turner

Conditional neural processes (CNPs; Garnelo et al., 2018a) are attractive meta-learning models which produce well-calibrated predictions and are trainable via a simple maximum likelihood procedure. Although CNPs have many advantages, they are unable to model dependencies in their predictions. Various works propose solutions to this, but these come at the cost of either requiring approximate inference or being limited to Gaussian predictions. In this work, we instead propose to change how CNPs are deployed at test time, without any modifications to the model or training procedure. Instead of making predictions independently for every target point, we autoregressively define a joint predictive distribution using the chain rule of probability, taking inspiration from the neural autoregressive density estimator (NADE) literature. We show that this simple procedure allows factorised Gaussian CNPs to model highly dependent, non-Gaussian predictive distributions. Perhaps surprisingly, in an extensive range of tasks with synthetic and real data, we show that CNPs in autoregressive (AR) mode not only significantly outperform non-AR CNPs, but are also competitive with more sophisticated models that are significantly more computationally expensive and challenging to train. This performance is remarkable given that AR CNPs are not trained to model joint dependencies. Our work provides an example of how ideas from neural distribution estimation can benefit neural processes, and motivates research into the AR deployment of other neural process models.

****

$k$NN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference

Benfeng Xu,Quan Wang,Zhendong Mao,Yajuan Lyu,Qiaoqiao She,Yongdong Zhang

In-Context Learning (ICL), which formulates target tasks as prompt completion conditioned on in-context demonstrations, has become the prevailing utilization of LLMs. In this paper, we first disclose an actual predicament for this typical usage that it can not scale up with training data due to context length restriction. Besides, existing works have shown that ICL also suffers from various biases and requires delicate calibration treatment. To address both challenges, we advocate a simple and effective solution, $k$NN Prompting, which first queries LLM with training data for distributed representations, then predicts test instances by simply referring to nearest neighbors. We conduct comprehensive experiments to demonstrate its two-fold superiority: 1) Calibration-Free: $k$NN Prompting does not directly align LLM output distribution with task-specific label space, instead leverages such distribution to align test and training instances. It significantly outperforms state-of-the-art calibration-based methods under comparable few-shot scenario. 2) Beyond-Context: $k$NN Prompting can further scale up effectively with as many training data as are available, continually bringing substantial improvements. The scaling trend holds across 10 orders of magnitude ranging from 2 shots to 1024 shots as well as different LLMs scales ranging from 0.8B to 30B. It successfully bridges data scaling into model scaling, and brings new potentials for the gradient-free paradigm of LLM deployment. Code is publicly available at https://github.com/BenfengXu/KNNPrompting

****

Closed-loop Transcription via Convolutional Sparse Coding

Xili Dai,Ke Chen,Shengbang Tong,Jingyuan Zhang,Xingjian Gao,Yuexiang Zhai,Mingyang Li,Xiaojun Yuan,Heung-Yeung Shum,Lionel Ni,Yi Ma

Autoencoding has been a popular and effective framework for learning generative models for images, with much empirical success. Autoencoders often use generic d eep networks as the encoder and decoder, which are difficult to interpret, and the learned representations lack clear structure. In this work, we replace the encoder and decoder with standard convolutional sparse coding and decoding layers, obtained from unrolling an optimization algorithm for solving a (convexified) sparse coding program. Furthermore, to avoid computational difficulties in minimizing distributional distance between the real and generated images, we utilize the recent closed-loop transcription (CTRL) framework that maximizes the rate reduction of the learned sparse representations. We show that such a simple framework demonstrates surprisingly competitive performance on large datasets, such as ImageNet-1K, compared to existing autoencoding and generative methods under fair conditions. Even with simpler networks and less computational resources, our method demonstrates splendid visual quality in regenerated images with striking sample-wise consistency. More surprisingly, the learned autoencoder generalizes to unseen datasets. Our method enjoys several side benefits, including more structured and interpretable representations, more stable convergence, scalability to large datasets -- indeed, our method is the first sparse coding generative method to scale up to ImageNet -- and trainability with smaller batch sizes.

****

Transformers Learn Shortcuts to Automata

Bingbin Liu,Jordan T. Ash,Surbhi Goel,Akshay Krishnamurthy,Cyril Zhang

Algorithmic reasoning requires capabilities which are most naturally understood through recurrent models of computation, like the Turing machine. However, Transformer models, while lacking recurrence, are able to perform such reasoning using far fewer layers than the number of reasoning steps. This raises the question: what solutions are these shallow and non-recurrent models finding? We investigate this question in the setting of learning automata, discrete dynamical systems naturally suited to recurrent modeling and expressing algorithmic tasks. Our theoretical results completely characterize shortcut solutions, whereby a shallow Transformer with only $o(T)$ layers can exactly replicate the computation of an automaton on an input sequence of length T. By representing automata using the algebraic structure of their underlying transformation semigroups, we obtain $O(\log T)$-depth simulators for all automata and $O(1)$-depth simulators for all automata whose associated groups are solvable. Empirically, we perform synthetic experiments by training Transformers to simulate a wide variety of automata, and show that shortcut solutions can be learned via standard training. We further investigate the brittleness of these solutions and propose potential mitigations.

****

Efficient neural representation in the cognitive neuroscience domain: Manifold Capacity in One-vs-rest Recognition Limit

Nga Yu Lo,SueYeon Chung

The structure in neural representations as manifolds has become a popular approach to study information encoding in neural populations. One particular interest is the connection between object recognition capability and the separability of neural representations for different objects, often called "object manifolds." In learning theory, separability has been studied under the notion of storage capacity, which refers to the number of patterns encoded in a feature dimension. Chung et al (2018) extended the notion of capacity from discrete points to manifolds, where manifold capacity refers to the maximum number of object manifolds that can be linearly separated with high probability given random assignment of labels. Despite the use of manifold capacity in analyzing artificial neural networks (ANNs), its application to neuroscience has been limited. Due to the limited number of "features", such as neurons, available in neural experiments, manifold capacity cannot be verified empirically, unlike in ANNs. Additionally, the usage of random label assignment, while common in learning theory, is of limited relevance to the definition of object recognition tasks in cognitive science. To overcome these limits, we present the Sparse Replica Manifold analysis to study object recognition. Sparse manifold capacity measures how many object manifolds can be separated under one versus the rest classification, a form of task widely used in both in cognitive neuroscience experiments and machine learning applicatio ns. We demonstrate the application of sparse manifold capacity allows analysis of a wider class of neural data - in particular, neural data that has a limited number of neurons with empirical measurements. Furthermore, sparse manifold capacity requires less computations to evaluate underlying geometries and enables a connection to a measure of dimension, the participation ratio. We analyze the relationship between capacity and dimension, and demonstrate that both manifold intrinsic dimension and the ambient space dimension play a role in capacity.

****

ULF: UNSUPERVISED LABELING FUNCTION CORRECTION USING CROSS-VALIDATION FOR WEAK SUPERVISION

Anastasiia Sedova,Benjamin Roth

A way to overcome expensive and time-consuming manual data labeling is weak supervision - automatic annotation of data samples via a predefined set of labeling functions (LFs), rule-based mechanisms that generate artificial labels for the classes associated with the LFs. In this work, we investigate noise reduction techniques for weak supervision based on the principle of k-fold cross-validation. We introduce a new algorithm ULF for denoising weakly annotated data which uses models trained on all but some LFs to detect and correct biases specific to the held-out LFs. Specifically, ULF refines the allocation of LFs to classes by re-estimating this assignment on highly reliable cross-validated samples. We realize two variants of this algorithm: feature-based ULF (relying on count-based feature vectors), and DeepULF (fine-tuning pre-trained language models). We compare ULF to methods originally developed for detecting erroneous samples in manually annotated data, as well as to our extensions of such methods to the weakly supervised setting. Our new weak supervision-specific methods (ULF and extensions) leverage the information about matching LFs, making detecting noisy samples more accurate. Evaluation on several datasets shows that ULF can successfully improve weakly supervised learning without utilizing any manually labeled data.

****

Islands of Confidence: Robust Neural Network Classification with Uncertainty Quantification

Sibylle Hess,Tianjin Huang,Wouter Duivesteijn

We propose a Gaussian confidence measure and its optimization, for use in neural network classifiers. The measure comes with theoretical results, simultaneously resolving two pressing problems in NN classification: uncertainty quantification, and robustness. Existing research in uncertainty quantification mostly revolves around the confidence reflected in the input feature space. Instead, we focus on the learned representation of the network and analyze the confidence in the penultimate layer space. We formally prove that, independent of optimization-procedural effects, a set of centroids always exists such that softmax classifiers are nearest-centroid classifiers. Softmax confidence, however, does not reflect that the classification is based on nearest centroids: artificially inflated confidence is also given to out-of-distributions samples that are not near any centroid, but slightly less distant from one centroid than from the others. Our new confidence measure is centroid-based, and hence no longer suffers from the artificial confidence inflation of out-of-distribution samples. We also show that our proposed centroidal confidence measure is providing a robustness certificate against attacks. As such, it manages to reflect what the model doesn't know (as demanded by uncertainty quantification), and to resolve the issue of robustness of neural networks.

****

Quantization-aware Policy Distillation (QPD)

Thomas Avé,Kevin Mets,Tom De Schepper,Steven Latre

Recent advancements have made Deep Reinforcement Learning (DRL) exceedingly more powerful, but the produced models remain very computationally complex and therefore difficult to deploy on edge devices.

Compression methods such as quantization and distillation can be used to increase the applicability of DRL models on these low-power edge devices by decreasing the necessary precision and number of operations respectively.

Training in low-precision is notoriously less stable however, which is amplified by the decrease in representational power when limiting the number of trainable parameters.

We propose Quantization-aware Policy Distillation (QPD), which overcomes this instability by providing a smoother transition from high to low-precision network parameters.

A new distillation loss specifically designed for the compression of actor-critic networks is also defined, resulting in a higher accuracy after compression.

Our experiments show that these combined methods can effectively compress a network down to 0.5% of its original size, without any loss in performance.

****

FIFA: Making Fairness More Generalizable in Classifiers Trained on Imbalanced Data

Zhun Deng,Jiayao Zhang,Linjun Zhang,Ting Ye,Yates Coley,Weijie J Su,James Zou

Algorithmic fairness plays an important role in machine learning and imposing fairness constraints during learning is a common approach. However, many datasets are imbalanced in certain label classes (e.g. "healthy") and sensitive subgroups (e.g. "older patients"). Empirically, this imbalance leads to a lack of generalizability not only of classification but also of fairness properties, especially in over-parameterized models. For example, fairness-aware training may ensure equalized odds (EO) on the training data, but EO is far from being satisfied on new users. In this paper, we propose a theoretically-principled, yet {\bf F}lexible approach that is {\bf I}mbalance-{\bf F}airness-{\bf A}ware ({\bf FIFA}). Specifically, FIFA encourages both classification and fairness generalization and can be flexibly combined with many existing fair learning methods with logits-based losses. While our main focus is on EO, FIFA can be directly applied to achieve equalized opportunity (EqOpt); and under certain conditions, it can also be applied to other fairness notions. We demonstrate the power of FIFA by combining it with a popular fair classification algorithm, and the resulting algorithm achieves significantly better fairness generalization on several real-world datasets.

****

GMML is All you Need

Sara Atito,Muhammad Awais,Josef Kittler

Vision transformers have generated significant interest in the computer vision (CV) community because of their flexibility in exploiting contextual information, whether it is sharply confined local, or long range global. However, they are known to be data hungry. This has motivated the research in self-supervised transformer pretraining, which does not need to decode the semantic information conveyed by labels to link it to the image properties, but rather focuses directly on extracting a concise representation of the image data that reflects the notion of similarity and is invariant to nuisance factors. The key vehicle for the self-learning process used by the majority of self-learning methods is the generation of multiple views of the training data and the creation of pretext tasks which use these views to define the notion of image similarity and data integrity. However, this approach lacks the natural propensity to extract contextual information. We propose group mask model learning (GMML), a self-supervised learning (SSL) mechanism for pretraining vision transformers with the ability to extract the contextual information present in all the concepts in an image. GMML achieves this by manipulating random groups of connected tokens, ensuingly covering a meaningful part of a semantic concept, and then recovering the hidden semantic information from the visible part of the concept. GMML implicitly introduces a novel data augmentation process. Unlike most of the existing SSL approaches, GMML does not require momentum encoder, nor rely on careful implementation details such as large batches and gradient stopping, which are all artefacts of most of the current self-supervised learning techniques. Since its conception at the beginning of 2021, GMML maintains itself as unbeaten SSL method with several desirable benefits and marked a significant milestone in computer vision by being one of the first self-supervised pretraining methods which outperform supervised pretraining consistently with a large margin. GMML is simple, elegant, and currently the best mechanism to extract information from a given dataset and instil this infor mation into transformer's weights. The code will be made publicly available for the community to train on bigger corpora.

****

Understanding The Robustness of Self-supervised Learning Through Topic Modeling

Zeping Luo,Shiyou Wu,Cindy Weng,Mo Zhou,Rong Ge

Self-supervised learning has significantly improved the performance of many NLP tasks. However, how can self-supervised learning discover useful features, and why is it better than traditional approaches such as probabilistic models are still largely unknown. In this paper, we focus on the context of topic modeling and highlight a key advantage of self-supervised learning - when applied to data generated by topic models, self-supervised learning can be oblivious to the specific model, and hence is less susceptible to model misspecification. In particular, we prove that commonly used self-supervised objectives based on reconstruction or contrastive samples can both recover useful posterior information for general topic models. Empirically, we show that the same objectives can perform on par with posterior inference using the correct model, while outperforming posterior inference using misspecified models.

****

Temporal Disentanglement of Representations for Improved Generalisation in Reinforcement Learning

Mhairi Dunion,Trevor McInroe,Kevin Sebastian Luck,Josiah P. Hanna,Stefano V Albrecht

Reinforcement Learning (RL) agents are often unable to generalise well to environment variations in the state space that were not observed during training. This issue is especially problematic for image-based RL, where a change in just one variable, such as the background colour, can change many pixels in the image. The changed pixels can lead to drastic changes in the agent's latent representation of the image, causing the learned policy to fail. To learn more robust representations, we introduce TEmporal Disentanglement (TED), a self-supervised auxiliary task that leads to disentangled image representations exploiting the sequential nature of RL observations. We find empirically that RL algorithms utilising TED as an auxiliary task adapt more quickly to changes in environment variables with continued training compared to state-of-the-art representation learning methods. Since TED enforces a disentangled structure of the representation, our experiments also show that policies trained with TED generalise better to unseen values of variables irrelevant to the task (e.g. background colour) as well as unseen values of variables that affect the optimal policy (e.g. goal positions).

****

VIPeR: Provably Efficient Algorithm for Offline RL with Neural Function Approximation

Thanh Nguyen-Tang,Raman Arora

We propose a novel algorithm for offline reinforcement learning called Value Iteration with Perturbed Rewards (VIPeR), which amalgamates the pessimism principle with random perturbations of the value function. Most current offline RL algorithms explicitly construct statistical confidence regions to obtain pessimism via lower confidence bounds (LCB), which cannot easily scale to complex problems where a neural network is used to estimate the value functions. Instead, VIPeR implicitly obtains pessimism by simply perturbing the offline data multiple times with carefully-designed i.i.d. Gaussian noises to learn an ensemble of estimated state-action {value functions} and acting greedily with respect to the minimum of the ensemble. The estimated state-action values are obtained by fitting a parametric model (e.g., neural networks) to the perturbed datasets using gradient descent. As a result, VIPeR only needs $\mathcal{O}(1)$ time complexity for action selection, while LCB-based algorithms require at least $\Omega(K^2)$, where K is the total number of trajectories in the offline data. We also propose a novel data-splitting technique that helps remove a factor involving the log of the covering number in our bound. We prove that VIPeR yields a provable uncertainty quantifier with overparameterized neural networks and enjoys a bound on sub-optimality of $\tilde{\mathcal{O}}({ \kappa H^{5/2} \tilde{d} }/{\sqrt{K}})$, where $\tilde{d}$ is the effective dimension, H is the horizon length and κ

measures the distributional shift. We corroborate the statistical and computational efficiency of VIPeR with an empirical evaluation on a wide set of synthetic and real-world datasets. To the best of our knowledge, VIPeR is the first algorithm for offline RL that is provably efficient for general Markov decision processes (MDPs) with neural network function approximation.

****

Deep Probabilistic Time Series Forecasting over Long Horizons

Gregory Benton,Nate Gruver,Wesley Maddox,Andrew Gordon Wilson

Recent advances in neural network architectures for time series have led to significant improvements on deterministic forecasting metrics like mean squared error. We show that for many common benchmark datasets with deterministic evaluation metrics, intrinsic stochasticity is so significant that simply predicting summary statistics of the inputs outperforms many state-of-the-art methods, despite these simple forecasters capturing essentially no information from the noisy signals in the dataset. We demonstrate that using a probabilistic framework and moving away from deterministic evaluation acts as a simple fix for this apparent misalignment between good performance and poor understanding. With simple and scalable approaches for uncertainty representation we can adapt state-of-the-art architectures for point prediction to be excellent probabilistic forecasters, outperforming complex probabilistic methods constructed from deep generative models (DGMs) on popular benchmarks. Finally, we demonstrate that our simple adaptations to point predictors yield reliable probabilistic forecasts on many problems of practical significance, namely large and highly stochastic datasets of climatological and economic data.

****

Revealing Dominant Eigendirections via Spectral Non-Robustness Analysis in the Deep Reinforcement Learning Policy Manifold

Ezgi Korkmaz

Deep neural policies have recently been installed in a diverse set of settings, from biotechnology to automated financial systems. However, the utilization of deep neural networks to approximate the state-action value function commences concerns on the decision boundary stability, in particular, with regard to the sensitivity of policy decision making to indiscernible, non-robust features due to highly non-convex and complex deep neural manifolds. These concerns constitute an obstruction to understanding the reasoning made by deep neural policies, and their foundational limitations. Thus, it is crucial to develop techniques that aim to understand the sensitivities in the learnt representations of neural network policies. To achieve this we introduce a method that identifies the dominant eigen-directions via spectral analysis of non-robust directions in the deep neural policy decision boundary across both time and space. Through experiments in the Arcade Learning Environment (ALE), we demonstrate the effectiveness of our spectral analysis algorithm for identifying correlated non-robust directions, and for measuring how sample shifts remold the set of sensitive directions in the neural policy landscape. Most importantly, we show that state-of-the-art adversarial training techniques yield learning of sparser high-sensitivity directions, with dramatically larger oscillations over time, when compared to standard training. We believe our results reveal the fundamental properties of the decision process made by the deep reinforcement learning policies, and can help in constructing safe, reliable and value-aligned deep neural policies.

****

MC-SSL: Towards Multi-Concept Self-Supervised Learning

Sara Atito,Muhammad Awais,Ammarah Farooq,Zhenhua Feng,Josef Kittler

Self-supervised pre-training is the method of choice for natural language processing models and is rapidly gaining popularity in many vision tasks. Recently, self-supervised pre-training has shown to outperform supervised pre-training for many downstream vision applications, marking a milestone in the area. This superiority is attributed to the negative impact of incomplete labelling of the training images, which convey multiple concepts, but are annotated using a single dominant class label. Although Self-Supervised Learning (SSL), in principle, is free of this limitation, the choice of a pretext task facilitating SSL can perpetuat e this shortcoming by driving the learning process towards a single concept output. This study aims to investigate the possibility of modelling all the concepts present in an image without using labels. In this respect the proposed Multi-Concept SSL (MC-SSL) framework is a step towards unsupervised learning which embraces all the diverse content in an image with the aim of explicitly modelling the information from all the concepts present in the image. MC-SSL involves two core design steps: group masked model learning (GMML) and learning of pseudo-concepts for data tokens using a momentum encoder (teacher-student) framework. An added benefit of MC-SSL is the ability to train data hungry transformers on small datasets with high accuracy without external data. Experimental results on multi-label and multi-class image classification downstream tasks demonstrate that MC-SSL not only surpasses existing SSL methods but also outperforms supervised transfer learning. The source code will be made publicly available for the community to train on bigger corpus.

****

Latent Hierarchical Imitation Learning for Stochastic Environments

Maximilian Igl,Punit Shah,Paul Mougin,Sirish Srinivasan,Tarun Gupta,Brandyn White,Kyriacos Shiarlis,Shimon Whiteson

Many applications of imitation learning require the agent to avoid mode collapse and mirrorthe full distribution of observed behaviors. Existing methods improving this distributional realism typically rely on hierarchical policies conditioned on sampled types that model agent-internal features like persona, goal, or strategy. However, these methods are often inappropriate for stochastic environments, where internal and external factors of influence on the observed agent trajectories have to be disentangled, and only internal factors should be encoded in the agent type to be robust to changing environment conditions. We formalize this challenge as distribution shifts in the marginal and conditional distributions of agent types under environmental stochasticity, in addition to the familiar covariate shift in state visitations. We propose Robust Type Conditioning (RTC), which eliminates these shifts with adversarial training under randomly sampled types. Experiments on two domains, including the large-scal eWaymo Open Motion Dataset, show improved distributional realism while maintaining or improving task performance compared to state of the art baselines.

****

Exploring the Limits of Differentially Private Deep Learning with Group-wise Clipping

Jiyan He,Xuechen Li,Da Yu,Huishuai Zhang,Janardhan Kulkarni,Yin Tat Lee,Arturs Backurs,Nenghai Yu,Jiang Bian

Differentially private deep learning has recently witnessed advances in computational efficiency and privacy-utility trade-off. We explore whether further improvements along the two axes are possible and provide affirmative answers leveraging two instantiations of \emph{group-wise clipping}. To reduce the compute time overhead of private learning, we show that \emph{per-layer clipping}, where the gradient of each neural network layer is clipped separately, allows clipping to be performed in conjunction with backpropagation in differentially private optimization. This results in private learning that is as memory-efficient and almost as fast per training update as non-private learning for many workflows of interest. While per-layer clipping with constant thresholds tends to underperform standard flat clipping, per-layer clipping with adaptive thresholds matches or outperforms flat clipping under given training epoch constraints, hence attaining similar or better task performance within less wall time. To explore the limits of scaling (pretrained) models in differentially private deep learning, we privately fine-tune the 175 billion-parameter GPT-3. We bypass scaling challenges associated with clipping gradients that are distributed across multiple devices with \emph{per-device clipping} that clips the gradient of each model piece separately on its host device. Privately fine-tuning GPT-3 with per-device clipping achieves a task performance at $\epsilon=1$ better than what is attainable by non-privately fine-tuning the largest GPT-2 on a summarization task.

****

Mesh-free Eulerian Physics-Informed Neural Networks

Fabricio Arend Torres,Marcello Massimo Negri,Monika Nagy-Huber,Maxim Samarin,Volker Roth

Physics-informed Neural Networks (PINNs) have recently emerged as a principled way to include prior physical knowledge in form of partial differential equations (PDEs) into neural networks. Although PINNs are generally viewed as mesh-free, current approaches still rely on collocation points within a bounded region, even in settings with spatially sparse signals. Furthermore, if the boundaries are not known, the selection of such a region is difficult and often results in a large proportion of collocation points being selected in areas of low relevance. To resolve this severe drawback of current methods, we present a mesh-free and adaptive approach termed particle-density PINN (pdPINN), which is inspired by the microscopic viewpoint of fluid dynamics. The method is based on the Eulerian formulation and, different from classical mesh-free method, does not require the introduction of Lagrangian updates. We propose to sample directly from the distribution over the particle positions, eliminating the need to introduce boundaries while adaptively focusing on the most relevant regions. This is achieved by interpreting a non-negative physical quantity (such as the density or temperature) as an unnormalized probability distribution from which we sample with dynamic Monte Carlo methods. The proposed method leads to higher sample efficiency and improved performance of PINNs. These advantages are demonstrated on various experiments based on the continuity equations, Fokker-Planck equations, and the heat equation.

****

Self-supervised learning with rotation-invariant kernels

Léon Zheng,Gilles Puy,Elisa Riccietti,Patrick Perez,Rémi Gribonval

We introduce a regularization loss based on kernel mean embeddings with rotation-invariant kernels on the hypersphere (also known as dot-product kernels) for self-supervised learning of image representations. Besides being fully competitive with the state of the art, our method significantly reduces time and memory complexity for self-supervised training, making it implementable for very large embedding dimensions on existing devices and more easily adjustable than previous methods to settings with limited resources. Our work follows the major paradigm where the model learns to be invariant to some predefined image transformations (cropping, blurring, color jittering, etc.), while avoiding a degenerate solution by regularizing the embedding distribution. Our particular contribution is to propose a loss family promoting the embedding distribution to be close to the uniform distribution on the hypersphere, with respect to the maximum mean discrepancy pseudometric. We demonstrate that this family encompasses several regularizers of former methods, including uniformity-based and information-maximization methods, which are variants of our flexible regularization loss with different kernels. Beyond its practical consequences for state of the art self-supervised learning with limited resources, the proposed generic regularization approach opens perspectives to leverage more widely the literature on kernel methods in order to improve self-supervised learning methods.

****

Strong inductive biases provably prevent harmless interpolation

Michael Aerni,Marco Milanta,Konstantin Donhauser,Fanny Yang

Classical wisdom suggests that estimators should avoid fitting noise to achieve good generalization. In contrast, modern overparameterized models can yield small test error despite interpolating noise — a phenomenon often called "benign overfitting" or "harmless interpolation". This paper argues that the degree to which interpolation is harmless hinges upon the strength of an estimator's inductive bias, i.e., how heavily the estimator favors solutions with a certain structure: while strong inductive biases prevent harmless interpolation, weak inductive biases can even require fitting noise to generalize well. Our main theoretical result establishes tight non-asymptotic bounds for high-dimensional kernel regression that reflect this phenomenon for convolutional kernels, where the filter size regulates the strength of the inductive bias. We further provide empirical evidence of the same behavior for deep neural networks with varying filter sizes and rotational invariance.

****

Active Learning based Structural Inference

Aoran Wang,Jun Pang

In this paper, we propose an active-learning based framework, Active Learning based Structural Inference (ALaSI), to infer the existence of directed connections from observed agents' states over a time period in a dynamical system. With the help of deep active learning, ALaSI is competent in learning the representation of connections with relatively small pool of prior knowledge. Moreover, based on information theory, we propose inter- and out-of-scope message learning pipelines, which are remarkably beneficial to the structural inference for large dynamical systems. We evaluate ALaSI on various large datasets including simulated systems and real-world networks, to demonstrate that ALaSI is able to precisely infer the existence of connections in these systems under either supervised learning or unsupervised learning, with better performance than baseline methods.

****

Batch Normalization Explained

Randall Balestriero,Richard Baraniuk

A critically important, ubiquitous, and yet poorly understood ingredient in modern deep networks (DNs) is batch normalization (BN), which centers and normalizes the feature maps. To date, only limited progress has been made understanding why BN boosts DN learning and inference performance; work has focused exclusively on showing that BN smooths a DN's loss landscape. In this paper, we study BN theoretically from the perspective of function approximation; we exploit the fact that most of today's state-of-the-art DNs are continuous piecewise affine (CPA) splines that fit a predictor to the training data via affine mappings defined over a partition of the input space (the so-called ``linear regions''). We demonstrate that BN is an unsupervised learning technique that -- independent of the DN's weights or gradient-based learning -- adapts the geometry of a DN's spline partition to match the data. BN provides a ``smart initialization'' that boosts the performance of DN learning, because it adapts even a DN initialized with random weights to align its spline partition with the data. We also show that the variation of BN statistics between mini-batches introduces a dropout-like random perturbation to the partition boundaries and hence the decision boundary for classification problems. This per mini-batch perturbation reduces overfitting and improves generalization by increasing the margin between the training samples and the decision boundary.

****

AN OPERATOR NORM BASED PASSIVE FILTER PRUNING METHOD FOR EFFICIENT CNNS

Arshdeep Singh,Yunpeng Li,Mark D Plumbley

Convolutional neural networks (CNNs) have shown state-of-the-art performance in various applications. However, CNNs are resource-hungry due to their requirement of high computational complexity and memory storage. Recent efforts toward achieving computational efficiency in CNNs involve filter pruning methods that eliminate some of the filters in CNNs based on the "importance" of the filters. Existing passive filter pruning methods typically use the entry-wise norm of the filters to quantify filter importance, without considering how well the filter contributes in producing the node output. Under situations where the large number of filters are to be pruned from the network, the entry-wise norm methods always select high entry-wise norm filters as important, and ignore the diversity learned by the other filters that may result in degradation in the performance. To address this, we present a passive filter pruning method where the filters are pruned based on their contribution in producing output by implicitly considering the operator norm of the filters. The computational cost and memory requirement is reduced significantly by eliminating filters and their corresponding feature maps from the network. Accuracy similar to the original network is recovered by fine-tuning the pruned network. The proposed pruning method gives similar or better performance and recovers accuracy faster during the fine-tuning process than the entry-wise norm-based pruning methods. The efficacy of the proposed pruning method is evaluated on audio scene classification (e.g. TAU Urban Acoustic Scenes 2020) and image classification (MNIST handwritten digit classification).

****

Neuromechanical Autoencoders: Learning to Couple Elastic and Neural Network Nonlinearity

Deniz Oktay,Mehran Mirramezani,Eder Medina,Ryan P Adams

Intelligent biological systems are characterized by their embodiment in a complex environment and the intimate interplay between their nervous systems and the nonlinear mechanical properties of their bodies. This coordination, in which the dynamics of the motor system co-evolved to reduce the computational burden on the brain, is referred to as "mechanical intelligence" or "morphological computation". In this work, we seek to develop machine learning analogs of this process, in which we jointly learn the morphology of complex nonlinear elastic solids along with a deep neural network to control it. By using a specialized differentiable simulator of elastic mechanics coupled to conventional deep learning architectures---which we refer to as neuromechanical autoencoders---we are able to learn to perform morphological computation via gradient descent. Key to our approach is the use of mechanical metamaterials---cellular solids, in particular---as the morphological substrate. Just as deep neural networks provide flexible and massively-parametric function approximators for perceptual and control tasks, cellular solid metamaterials are promising as a rich and learnable space for approximating a variety of actuation tasks. In this work we take advantage of these complementary computational concepts to co-design materials and neural network controls to achieve nonintuitive mechanical behavior. We demonstrate in simulation how it is possible to achieve translation, rotation, and shape matching, as well as a "digital MNIST" task. We additionally manufacture and evaluate one of the designs to verify its real-world behavior.

****

Temporal Dynamics Aware Adversarial Attacks On Discrete-Time Graph Models

Kartik Sharma,Rakshit Trivedi,Rohit Sridhar,Srijan Kumar

Real-world graphs such as social networks, communication networks, and rating networks are constantly evolving over time. Many architectures have been developed to learn effective node representations using both graph structure and its dynamics. While the robustness of static graph models is well-studied, the vulnerability of the dynamic graph models to adversarial attacks is underexplored. In this work, we design a novel adversarial attack on discrete-time dynamic graph models where we desire to perturb the input graph sequence in a manner that preserves the temporal dynamics of the graph. To this end, we motivate a novel Temporal Dynamics-Aware Perturbation (TDAP) constraint, which ensures that perturbations introduced at each time step are restricted to only a small fraction of the number of changes in the graph since the previous time step. We present a theoretically-grounded Projected Gradient Descent approach for dynamic graphs to find the effective perturbations under the TDAP constraint. Experiments on two tasks — dynamic link prediction and node classification, show that our approach is up to 4x more effective than the baseline methods for attacking these models. We also consider the practical online setting where graph snapshots become available in real-time and extend our attack approach to use Online Gradient Descent for performing attacks under the TDAP constraint. In this more challenging setting, we demonstrate that our method achieves upto 5x superior performance when compared to representative baselines.

****

Automatic Curriculum Generation for Reinforcement Learning in Zero-Sum Games

Jiayu Chen,Yunfei Li,Zelai Xu,Huazhong Yang,Jiaming Song,Yu Wang,Yi Wu

Curriculum learning (CL), whose core idea is to train from easy to hard, is a popular technique to accelerate reinforcement learning (RL) training. It has also been a trend to automate the curriculum generation process. Automatic CL works primarily focus on goal-conditioned RL problems, where an explicit indicator of training progress, e.g., reward or success rate, can be used to prioritize the training tasks. However, such a requirement is no longer valid in zero-sum games: there are no goals for the agents, and the accumulative reward of the learning policy can constantly fluctuate throughout training. In this work, we present the first theoretical framework of automatic curriculum learning in the setting of zero-sum games and derive a surprisingly simple indicator of training progress, i.e., the Q-value variance, which can be directly approximated by computing the variance of value network ensembles. With such a progression metric, we further adopt a particle-based task sampler to generate initial environment configurations for training, which is particularly lightweight, computation-efficient, and naturally multi-modal. Combining these techniques with multi-agent PPO training, we obtain our final algorithm, Zero-sum Automatic Curriculum Learning (ZACL). We first evaluate ZACL in a 2D particle-world environment, where ZACL produces much stronger policies than popular RL methods for zero-sum games using the same amount of samples. Then we show in the challenging hide-and-seek environment that ZACL can lead to all four emergent phases using a single desktop computer, which is reported for the first time in the literature. The project website is at https://sites.google.com/view/zacl.

****

Internet-augmented language models through few-shot prompting for open-domain question answering

Angeliki Lazaridou,Elena Gribovskaya,Wojciech Jan Stokowiec,Nikolai Grigorev

In this work, we aim to capitalize on the unique few-shot capabilities of large-scale language models (LSLMs) to overcome some of their challenges with respect to grounding to factual and up-to-date information. Motivated by semi-parametric lan4 guage models (LMs), which ground their decisions in external retrieved evidence, we use few-shot prompting to learn to condition LMs on information returned from the web using Google Search, a broad and constantly updated knowledge source. Our approach does not involve fine-tuning or learning additional parameters, thus making it applicable to any LM, offering therefore a strong baseline. Indeed, we find that LMs conditioned on the web surpass performance of closed-book models of similar, or even larger, model sizes in open-domain question answering. Finally, we find that increasing the inference-time compute of models, achieved via using multiple retrieved evidences to generate multiple answers followed by a reranking stage that uses scores generated by the same LMs, leads to better performance and alleviates lower performance of smaller few-shot LMs. All in all, our findings suggest that it might be beneficial to slow down the race towards the biggest model and instead shift attention towards finding more effective ways to use models, including but not limited to, better prompting or increasing inference-time compute.

****

VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training

Yecheng Jason Ma,Shagun Sodhani,Dinesh Jayaraman,Osbert Bastani,Vikash Kumar,Amy Zhang

Reward and representation learning are two long-standing challenges for learning an expanding set of robot manipulation skills from sensory observations. Given the inherent cost and scarcity of in-domain, task-specific robot data, learning from large, diverse, offline human videos has emerged as a promising path towards acquiring a generally useful visual representation for control; however, how these human videos can be used for general-purpose reward learning remains an open question. We introduce $\textbf{V}$alue-$\textbf{I}$mplicit $\textbf{P}$re-training (VIP), a self-supervised pre-trained visual representation capable of generating dense and smooth reward functions for unseen robotic tasks. VIP casts representation learning from human videos as an offline goal-conditioned reinforcement learning problem and derives a self-supervised dual goal-conditioned value-function objective that does not depend on actions, enabling pre-training on unlabeled human videos. Theoretically, VIP can be understood as a novel implicit time contrastive objective that generates a temporally smooth embedding, enabling the value function to be implicitly defined via the embedding distance, which can then be used to construct the reward for any goal-image specified downstream task. Trained on large-scale Ego4D human videos and without any fine-tuning on in-domain, task-specific data, VIP can provide dense visual reward for an extensive set of simulated and $\textbf{real-robot}$ tasks, enabling diverse reward-based

visual control methods and significantly outperforming all prior pre-trained representations. Notably, VIP can enable simple, few-shot offline RL on a suite of real-world robot tasks with as few as 20 trajectories.
****

Language Modeling Using Tensor Trains

Zhan Su,Yuqin Zhou,Benyou Wang,Qiuchi Li,Jakob Grue Simonsen

Tensor networks have previously been shown to have potential in language modelling in theory, but lack of practical evidence support. We propose a novel Tensor Train Language Model (TTLM) based on Tensor-Train decomposition. We prove that TTLM generalizes Second-order Recurrent Neural Networks (RNNs), Recurrent Arithmetic Circuits and Multiplicative Integration RNNs in the sense that the architecture of all of these are, essentially, special cases of that of TTLM. To show the usefulness of TTLM, we perform a principled experimental evaluation on language modeling tasks, showing that our proposed variants, TTLM-large and TTLM-Tiny, can be more effective than Vanilla RNN while TTLM-Tiny has the half of the model size.
****

Bridging the Gap to Real-World Object-Centric Learning

Maximilian Seitzer,Max Horn,Andrii Zadaianchuk,Dominik Zietlow,Tianjun Xiao,Carl-Johann Simon-Gabriel,Tong He,Zheng Zhang,Bernhard Schölkopf,Thomas Brox,Francesco Locatello

Humans naturally decompose their environment into entities at the appropriate level of abstraction to act in the world. Allowing machine learning algorithms to derive this decomposition in an unsupervised way has become an important line of research. However, current methods are restricted to simulated data or require additional information in the form of motion or depth in order to successfully discover objects. In this work, we overcome this limitation by showing that reconstructing features from models trained in a self-supervised manner is a sufficient training signal for object-centric representations to arise in a fully unsupervised way. Our approach, DINOSAUR, significantly out-performs existing object-centric learning models on simulated data and is the first unsupervised object-centric model that scales to real world-datasets such as COCO and PASCAL VOC. DINOSAUR is conceptually simple and shows competitive performance compared to more involved pipelines from the computer vision literature.
****

Towards a Unified Theoretical Understanding of Non-contrastive Learning via Rank Differential Mechanism

Zhijian Zhuo,Yifei Wang,Jinwen Ma,Yisen Wang

Recently, a variety of methods under the name of non-contrastive learning (like BYOL, SimSiam, SwAV, DINO) show that when equipped with some asymmetric architectural designs, aligning positive pairs alone is sufficient to attain good performance in self-supervised visual learning. Despite some understandings of some specific modules (like the predictor in BYOL), there is yet no unified theoretical understanding of how these seemingly different asymmetric designs can all avoid feature collapse, particularly considering methods that also work without the predictor (like DINO). In this work, we propose a unified theoretical understanding for existing variants of non-contrastive learning. Our theory named Rank Differential Mechanism (RDM) shows that all these asymmetric designs create a consistent rank difference in their dual-branch output features. This rank difference will provably lead to an improvement of effective dimensionality and alleviate either complete or dimensional feature collapse. Different from previous theories, our RDM theory is applicable to different asymmetric designs (with and without the predictor), and thus can serve as a unified understanding of existing non-contrastive learning methods. Besides, our RDM theory also provides practical guidelines for designing many new non-contrastive variants. We show that these variants indeed achieve comparable performance to existing methods on benchmark datasets, and some of them even outperform the baselines. Our code is available at \url{https://github.com/PKU-ML/Rank-Differential-Mechanism}.
****

Stay Moral and Explore: Learn to Behave Morally in Text-based Games

Zijing Shi,Meng Fang,Yunqiu Xu,Ling Chen,Yali Du

Reinforcement learning (RL) in text-based games has developed rapidly and achieved promising results. However, little effort has been expended to design agents that pursue objectives while behaving morally, which is a critical issue in the field of autonomous agents. In this paper, we propose a general framework named Moral Awareness Adaptive Learning (MorAL) that enhances the morality capacity of an agent using a plugin moral-aware learning model. The framework allows the agent to execute task learning and morality learning adaptively. The agent selects trajectories from past experiences during task learning. Meanwhile, the trajectories are used to conduct self-imitation learning with a moral-enhanced objective. In order to achieve the trade-off between morality and task progress, the agent uses the combination of task policy and moral policy for action selection. We evaluate on the Jiminy Cricket benchmark, a set of text-based games with various scenes and dense morality annotations. Our experiments demonstrate that, compared with strong contemporary value alignment approaches, the proposed framework improves task performance while reducing immoral behaviours in various games.

****

Efficient Discovery of Dynamical Laws in Symbolic Form

Sören Becker,Michal Klein,Alexander Neitz,Giambattista Parascandolo,Niki Kilbertus

We propose a transformer-based sequence-to-sequence model that recovers scalar ordinary differential equations (ODEs) in symbolic form from time-series data of a single observed solution trajectory of the ODE. Our method is efficiently scalable: after one-time pretraining on a large set of ODEs, we can infer the governing laws of a new observed solution in a few forward passes of the model. First, we generate and make available a large dataset of more than 3M ODEs together with more than 63M numerical solutions for different initial conditions that may serve as a useful benchmark for future work on machine learning for dynamical systems. Then we show that our model performs better or on par with existing methods in various test cases in terms of accurate symbolic recovery of the ODE, especially for more complex expressions. Reliably recovering the symbolic form of dynamical laws is important as it allows for further dissemination of the inferred dynamics as well as meaningful modifications for predictions under interventions.

****

Brain2GAN; Reconstructing perceived faces from the primate brain via StyleGAN3

Thirza Dado,Paolo Papale,Antonio Lozano,Lynn Le,Feng Wang,Marcel van Gerven,Pieter R. Roelfsema,Ya■mur Güçlütürk,Umut Güçlü

Neural coding characterizes the relationship between stimuli and their corresponding neural responses. The usage of synthesized yet photorealistic reality by generative adversarial networks (GANs) allows for superior control over these data: the underlying feature representations that account for the semantics in synthesized data are known a priori and their relationship is perfect rather than approximated post-hoc by feature extraction models. We exploit this property in neural decoding of multi-unit activity responses that we recorded from the primate brain upon presentation with synthesized face images in a passive fixation experiment. The face reconstructions we acquired from brain activity were astonishingly similar to the originally perceived face stimuli. This provides strong evidence that the neural face manifold and the disentangled w-latent space conditioned on StyleGAN3 (rather than the z-latent space of arbitrary GANs or other feature representations we encountered so far) share how they represent the high-level semantics of the high-dimensional space of faces.

****

Optimistic Exploration with Learned Features Provably Solves Markov Decision Processes with Neural Dynamics

Sirui Zheng,Lingxiao Wang,Shuang Qiu,Zuyue Fu,Zhuoran Yang,Csaba Szepesvari,Zhaoran Wang

Incorporated with the recent advances in deep learning, deep reinforcement learning (DRL) has achieved tremendous success in empirical study. However, analyzing DRL is still challenging due to the complexity of the neural network class. In

this paper, we address such a challenge by analyzing the Markov decision process (MDP) with neural dynamics, which covers several existing models as special cases, including the kernelized nonlinear regulator (KNR) model and the linear MDP. We propose a novel algorithm that designs exploration incentives via learnable representations of the dynamics model by embedding the neural dynamics into a kernel space induced by the system noise. We further establish an upper bound on the sample complexity of the algorithm, which demonstrates the sample efficiency of the algorithm. We highlight that, unlike previous analyses of RL algorithms with function approximation, our bound on the sample complexity does not depend on the Eluder dimension of the neural network class, which is known to be exponentially large (Dong et al., 2021).
****

Would decentralization hurt generalization?
Tongtian Zhu,Fengxiang He,Kaixuan Chen,Mingli Song,Dacheng Tao
Decentralized stochastic gradient descent (D-SGD) allows collaborative learning on massive devices without controlling of a central server. Existing theory suggests that the decentralization degrades the generalizability, which conflicts with experimental results in the large-batch settings that D-SGD generalize better than centralized SGD (C-SGD). This work presents new theory that reconciles the conflict between the two perspectives. We prove that D-SGD introduces an implicit regularization that simultaneously penalizes (1) the sharpness of the learned minima and (2) the consensus distance between the consensus model and local models. We then prove that the implicit regularization is amplified in the large-batch settings when the linear scaling rule is applied. We further analyze the escaping efficiency of D-SGD, which suggests that D-SGD favors super-quadratic flat minima. Experiments are in full agreement with our theory. The code will be released publicly. To our best knowledge, this is the first work on the implicit regularization and escaping efficiency of D-SGD.
****

Variational Pseudo Labels for Meta Test-time Adaptation
Sameer Ambekar,Zehao Xiao,Jiayi Shen,Xiantong Zhen,Cees G. M. Snoek
Test-time model adaptation has shown great effectiveness in generalizing over domain shifts. A most successful tactic for test-time adaptation conducts further optimization on the target data using the predictions by the source-trained model. However, due to domain shifts, the source-trained model predictions themselves can be largely inaccurate, which results in a model misspecified to the target data and therefore damages their adaptation ability. In this paper, we address test-time adaptation from a probabilistic perspective. We formulate model adaption as a probabilistic inference problem, which incorporates the uncertainty into source model predictions by modeling pseudo labels as distributions. Based on the probabilistic formalism, we propose variational pseudo labels that explore the information of neighboring target samples to improve pseudo labels and achieve a model better specified to target data. By a meta-learning paradigm, we train our model by simulating domain shifts and the test-time adaptation procedure. In doing so, our model learns the ability to generate more accurate pseudo-label distributions and to adapt to new domains. Experiments on three widely used datasets demonstrate the effectiveness of our proposal.
****

No-Regret Learning in Strongly Monotone Games Converges to a Nash Equilibrium
Zifan Wang,Yi Shen,Michael Zavlanos,Karl Henrik Johansson
This paper studies a class of online games involving multiple agents with continuous actions that aim to minimize their local loss functions. An open question in the study of online games is whether no-regret learning for such agents leads to a Nash equilibrium. We address this question by providing a sufficient condition for strongly monotone games that guarantees Nash equilibrium convergence in a time average sense. Furthermore, we show that the class of games for which no-regret learning leads to a Nash equilibrium can be expanded if some further information on the learning algorithm is known. Specifically, we provide relaxed sufficient conditions for first-order and zeroth-order gradient descent algorithms as well as for best response algorithms in which agents choose actions that best

respond to other players' actions during the last episode. We analyze the convergence rate for these algorithms and present numerical experiments on three economic market problems to illustrate their performance.

****

Generalized Belief Transport

Junqi Wang,PEI WANG,Patrick Shafto

Human learners have ability to adopt appropriate learning approaches depending on constraints such as prior on the hypothesis and urgency of decision. However, existing learning models are typically considered individually rather than in relation to one and other. To build agents that have the ability to move between different modes of learning over time, it is important to understand how learning models are related as points in a broader space of possibilities. We introduce a mathematical framework, Generalized Belief Transport (GBT), that unifies and generalizes prior models, including Bayesian inference, cooperative communication and classification, as parameterizations of three learning constraints within Unbalanced Optimal Transport (UOT). We visualize the space of learning models encoded by GBT as a cube which includes classic learning models as special points. We derive critical properties of this parameterized space including proving continuity and differentiability which is the basis for model interpolation, and study limiting behavior of the parameters, which allows attaching learning models on the boundaries. Moreover, we investigate the long-run behavior of GBT, explore convergence properties of models in GBT mathematical and computationally, and formulate conjectures about general behavior. We conclude with open questions and implications for more unified models of learning.

****

Adversarial Cheap Talk

Chris Lu,Timon Willi,Alistair Letcher,Jakob Nicolaus Foerster

Adversarial attacks in reinforcement learning (RL) often assume highly-privileged access to the victim’s parameters, environment, or data. Instead, this paper proposes a novel adversarial setting called a Cheap Talk MDP in which an Adversary can merely append deterministic messages to the Victim’s observation, resulting in a minimal range of influence. The Adversary cannot occlude ground truth, influence underlying environment dynamics or reward signals, introduce non-stationarity, add stochasticity, see the Victim’s actions, or access their parameters. Additionally, we present a simple meta-learning algorithm called Adversarial Cheap Talk (ACT) to train Adversaries in this setting. We demonstrate that an Adversary trained with ACT can still significantly influence the Victim’s training and testing performance, despite the highly constrained setting. Affecting train-time performance reveals a new attack vector and provides insight into the success and failure modes of existing RL algorithms. More specifically, we show that an ACT Adversary is capable of harming performance by interfering with the learner’s function approximation, or instead helping the Victim’s performance by outputting useful features. Finally, we show that an ACT Adversary can manipulate messages during train-time to directly and arbitrarily control the Victim at test-time.

****

Learning to Induce Causal Structure

Nan Rosemary Ke,Silvia Chiappa,Jane X Wang,Jorg Bornschein,Anirudh Goyal,Melanie Rey,Theophane Weber,Matthew Botvinick,Michael Curtis Mozer,Danilo Jimenez Rezende

The fundamental challenge in causal induction is to infer the underlying graph structure given observational and/or interventional data. Most existing causal induction algorithms operate by generating candidate graphs and evaluating them using either score-based methods (including continuous optimization) or independence tests. In our work, we instead treat the inference process as a black box and design a neural network architecture that learns the mapping from both observational and interventional data to graph structures via supervised training on synthetic graphs. The learned model generalizes to new synthetic graphs, is robust to train-test distribution shifts, and achieves state-of-the-art performance on naturalistic graphs for low sample complexity.

****

Personalized federated composite learning with forward-backward envelopes

Jiang Hu,Jinhyun Ahn,Na Li,Quanzheng Li

Federated composite optimization (FCO) is an optimization problem in federated learning whose loss function contains a non-smooth regularizer. It arises naturally in the applications of federated learning (FL) that involve requirements such as sparsity, low rankness, and monotonicity. In this study, we propose a personalization method, called pFedFBE, for FCO by using forward-backward envelope (FBE) as clients’ loss functions. With FBE, we not only decouple the personalized model from the global model, but also allow personalized models to be smooth and easily optimized. In spite of the nonsmoothness of FCO, pFedFBE shows the same convergence complexity results as FedAvg for FL with unconstrained smooth objectives. Numerical experiments are shown to demonstrate the effectiveness of our proposed method.

****

Tackling Imbalanced Class in Federated Learning via Class Distribution Estimation

You-Ru Lu,Xiaoqian Wang,Dengfeng Sun

Federated Learning (FL) has become an upsurging machine learning method due to its applicability in large-scale distributed system and its privacy-preserving property. However, in real-world applications, the presence of class imbalance issue, especially the mismatch between local and global class distribution, greatly degrades the performance of FL. Moreover, due to the privacy constrain, the class distribution information of clients can not be accessed directly. To tackle class imbalance issue under FL setting, a novel algorithm, FedRE, is proposed in this paper. We propose a new class distribution estimation method for the FedRE algorithm, which requires no extra client data information and thus has no privacy concern. Both experimental results and theoretical analysis are provided to support the validity of our distribution estimation method. The proposed algorithm is verified with several experiment, including different datasets with the presence of class imbalance and local-global distribution mismatch. The experimental results show that FedRE is effective and it outperforms other related methods in terms of both overall and minority class classification accuracy.

****

Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning

Zhendong Wang,Jonathan J Hunt,Mingyuan Zhou

Offline reinforcement learning (RL), which aims to learn an optimal policy using a previously collected static dataset, is an important paradigm of RL. Standard RL methods often perform poorly in this regime due to the function approximation errors on out-of-distribution actions. While a variety of regularization methods have been proposed to mitigate this issue, they are often constrained by policy classes with limited expressiveness that can lead to highly suboptimal solutions. In this paper, we propose representing the policy as a diffusion model, a recent class of highly-expressive deep generative models. We introduce Diffusion Q-learning (Diffusion-QL) that utilizes a conditional diffusion model to represent the policy. In our approach, we learn an action-value function and we add a term maximizing action-values into the training loss of the conditional diffusion model, which results in a loss that seeks optimal actions that are near the behavior policy. We show the expressiveness of the diffusion model-based policy, and the coupling of the behavior cloning and policy improvement under the diffusion model both contribute to the outstanding performance of Diffusion-QL. We illustrate the superiority of our method compared to prior works in a simple 2D bandit example with a multimodal behavior policy. We then show that our method can achieve state-of-the-art performance on the majority of the D4RL benchmark tasks.

****

Subquadratic Algorithms for Kernel Matrices via Kernel Density Estimation

Ainesh Bakshi,Piotr Indyk,Praneeth Kacham,Sandeep Silwal,Samson Zhou

Kernel matrices, as well as weighted graphs represented by them, are ubiquitous objects in machine learning, statistics and other related fields. The main drawback of using kernel methods (learning and inference using kernel matrices) is efficiency -- given n input points, most kernel-based algorithms need to materialize the full $n \times n$ kernel matrix before performing any subsequent computation, thus incurring $\Omega(n^2)$ runtime. Breaking this quadratic barrier for various problems has therefore, been a subject of extensive research efforts.

We break the quadratic barrier and obtain \emph{subquadratic} time algorithms for several fundamental linear-algebraic and graph processing primitives, including approximating the top eigenvalue and eigenvector, spectral sparsification, solving linear systems, local clustering, low-rank approximation, arboricity estimation and counting weighted triangles. We build on the recently developed Kernel Density Estimation framework, which (after preprocessing in time subquadratic in n) can return estimates of row/column sums of the kernel matrix. In particular, we develop efficient reductions from \emph{weighted vertex} and \emph{weighted edge sampling} on kernel graphs, \emph{simulating random walks} on kernel graphs, and \emph{importance sampling} on matrices to Kernel Density Estimation and show that we can generate samples from these distributions in \emph{sublinear} (in the support of the distribution) time. Our reductions are the central ingredient in each of our applications and we believe they may be of independent interest. We empirically demonstrate the efficacy of our algorithms on low-rank approximation (LRA) and spectral sparsification, where we observe a $\textbf{9x}$ decrease in the number of kernel evaluations over baselines for LRA and a $\textbf{41x}$ reduction in the graph size for spectral sparsification.
****

CASA: Bridging the Gap between Policy Improvement and Policy Evaluation with Conflict Averse Policy Iteration

Changnan Xiao,Haosen Shi,Jiajun Fan,Shihong Deng,Haiyan Yin

We study the problem of model-free reinforcement learning, which is often solved following the principle of Generalized Policy Iteration (GPI). While GPI is typically an interplay between policy evaluation and policy improvement, most conventional model-free methods with function approximation assume the independence of GPI steps, despite of the inherent connections between them. In this paper, we present a method that attempts to eliminate the inconsistency between policy evaluation step and policy improvement step, leading to a conflict averse GPI solution with gradient-based functional approximation. Our method is capital to balancing exploitation and exploration between policy-based and value-based methods and is applicable to existed policy-based and value-based methods. We conduct extensive experiments to study theoretical properties of our method and demonstrate the effectiveness of our method on Atari 200M benchmark.
****

Achieving Near-Optimal Individual Regret & Low Communications in Multi-Agent Bandits

Xuchuang Wang,Lin Yang,Yu-Zhen Janice Chen,Xutong Liu,Mohammad Hajiesmaili,Don Towsley,John C.S. Lui

Cooperative multi-agent multi-armed bandits (CM2AB) study how distributed agents cooperatively play the same multi-armed bandit game. Most existing CM2AB works focused on maximizing the group performance of all agents---the accumulation of all agents' individual performance (i.e., individual reward). However, in many applications, the performance of the system is more sensitive to the ``bad'' agent---the agent with the worst individual performance. For example, in a drone swarm, a ``bad'' agent may crash into other drones and severely degrade the system performance. In that case, the key of the learning algorithm design is to coordinate computational and communicational resources among agents so to optimize the individual learning performance of the ``bad'' agent. In CM2AB, maximizing the group performance is equivalent to minimizing the group regret of all agents, and minimizing the individual performance can be measured by minimizing the maximum (worst) individual regret among agents. Minimizing the maximum individual regret was largely ignored in prior literature, and currently, there is little work on how to minimize this objective with a low communication overhead. In this pap

er, we propose a near-optimal algorithm on both individual and group regrets, in addition, we also propose a novel communication module in the algorithm, which only needs \(O(\log (\log T))\) communication times where \(T\) is the number of decision rounds. We also conduct simulations to illustrate the advantage of our algorithm by comparing it to other known baselines.
****
Online Boundary-Free Continual Learning by Scheduled Data Prior
Hyunseo Koh,Minhyuk Seo,Jihwan Bang,Hwanjun Song,Deokki Hong,Seulki Park,Jung-Woo Ha,Jonghyun Choi
Typical continual learning setup assumes that the dataset is split into multiple discrete tasks. We argue that it is less realistic as the streamed data would have no notion of task boundary in real-world data. Here, we take a step forward to investigate more realistic online continual learning – learning continuously changing data distribution without explicit task boundary, which we call boundary-free setup. As there is no clear boundary of tasks, it is not obvious when and what information in the past to be preserved as a better remedy for the stability-plasticity dilemma. To this end, we propose a scheduled transfer of previously learned knowledge. We further propose a data-driven balancing between the knowledge in the past and the present in learning objective. Moreover, since it is not straight-forward to use the previously proposed forgetting measure without task boundaries, we further propose a novel forgetting measure based on information theory that can capture forgetting. We empirically evaluate our method on a Gaussian data stream, its periodic extension, which assumes periodic data distribution frequently observed in real-life data, as well as the conventional disjoint task-split. Our method outperforms prior arts by large margins in various setups, using four popular benchmark datasets – CIFAR-10, CIFAR-100, TinyImageNet and ImageNet.
****
HypeR: Multitask Hyper-Prompted Training Enables Large-Scale Retrieval Generalization
ZeFeng Cai,Chongyang Tao,Tao Shen,Can Xu,Xiubo Geng,Xin Alex Lin,Liang He,Daxin Jiang
Recently, large-scale text retrieval has made impressive progress, facilitating both information retrieval and downstream knowledge-intensive tasks (e.g., open-domain QA and dialogue). With a moderate amount of data, a neural text retriever can outperform traditional methods such as BM25 by a large step. However, while being applied to out-of-domain data, the performance of a neural retriever degrades considerably. Therefore, how to enable a retriever to perform more robustly across different domains or tasks and even show strong zero-shot transfer ability is critical for building scalable IR systems. To this end, we propose HypeR, a hyper-prompted training mechanism to enable uniform retrieval across tasks of different domains. Specifically, our approach jointly trains the query encoder with a shared prompt-based parameter pool and a prompt synthesizer that dynamically composes hyper-prompt for encoding each query from different tasks or domains. Besides, to avoid the mode collapse of prompt attention distribution for different queries, we design a contrastive prompt regularization that promotes the mode of prompt attention to be aligned and uniform. Through multi-task hyper-prompted training, our retriever can master the ability to dynamically represent different types of queries and transfer knowledge across different domains and tasks. Extensive experiments show our model attains better retrieval performance across different tasks and better zero-shot transfer ability compared with various previous methods.
****
Learning Rationalizable Equilibria in Multiplayer Games
Yuanhao Wang,Dingwen Kong,Yu Bai,Chi Jin
A natural goal in multi-agent learning is to learn \emph{rationalizable} behavior, where players learn to avoid any Iteratively Dominated Action (IDA). However, standard no-regret based equilibria-finding algorithms could take exponential samples to find such rationalizable strategies. In this paper, we first propose a simple yet sample-efficient algorithm for finding a rationalizable action profi

le in multi-player general-sum games under bandit feedback, which substantially improves over the results of Wu et al. We further develop algorithms with the first efficient guarantees for learning rationalizable Coarse Correlated Equilibria (CCE) and Correlated Equilibria (CE). Our algorithms incorporate several novel techniques to guarantee the elimination of IDA and no (swap-)regret simultaneously, including a correlated exploration scheme and adaptive learning rates, which may be of independent interest. We complement our results with a sample complexity lower bound showing the sharpness of our guarantees.

****

A Higher Precision Algorithm for Computing the 1-Wasserstein Distance

Pankaj K Agarwal,Sharath Raghvendra,Pouyan Shirzadian,Rachita Sowle

We consider the problem of computing the 1-Wasserstein distance $\mathcal{W}(\mu,\nu)$ between two d-dimensional discrete distributions μ and ν whose support lie within the unit hypercube. There are several algorithms that estimate $\mathcal{W}(\mu,\nu)$ within an additive error of ε. However, when $\mathcal{W}(\mu,\nu)$ is small, the additive error ε dominates, leading to noisy results. Consider any additive approximation algorithm with execution time $T(n,\varepsilon)$. We propose an algorithm that runs in $O(T(n,\varepsilon/d) \log n)$ time and boosts the accuracy of estimating $\mathcal{W}(\mu,\nu)$ from ε to an expected additive error of $\min\{\varepsilon, (d\log_{\sqrt{d}/\varepsilon} n)\mathcal{W}(\mu,\nu)\}$. For the special case where every point in the support of μ and ν has a mass of $1/n$ (also called the Euclidean Bipartite Matching problem), we describe an algorithm to boost the accuracy of any additive approximation algorithm from ε to an expected additive error of $\min\{\varepsilon, (d\log\log n)\mathcal{W}(\mu,\nu)\}$ in $O(T(n, \varepsilon/d)\log\log n)$ time.

****

Energy-Based Test Sample Adaptation for Domain Generalization

Zehao Xiao,Xiantong Zhen,Shengcai Liao,Cees G. M. Snoek

In this paper, we propose energy-based sample adaptation at test time for domain generalization. Where previous works adapt their models to target domains, we adapt the unseen target samples to source-trained models. To this end, we design a discriminative energy-based model, which is trained on source domains to jointly model the conditional distribution for classification and data distribution for sample adaptation. The model is optimized to simultaneously learn a classifier and an energy function. To adapt target samples to source distributions, we iteratively update the samples by energy minimization with stochastic gradient Langevin dynamics. Moreover, to preserve the categorical information in the sample during adaptation, we introduce a categorical latent variable into the energy-based model. The latent variable is learned from the original sample before adaptation by variational inference and fixed as a condition to guide the sample update. Experiments on six benchmarks for classification of images and microblog threads demonstrate the effectiveness of our proposal.

****

Representation Power of Graph Convolutions : Neural Tangent Kernel Analysis

Mahalakshmi Sabanayagam,Pascal Esser,Debarghya Ghoshdastidar

The fundamental principle of Graph Neural Networks (GNNs) is to exploit the structural information of the data by aggregating the neighboring nodes using a `graph convolution’. Therefore, understanding its influence on the network performance is crucial. Convolutions based on graph Laplacian have emerged as the dominant choice with the symmetric normalization of the adjacency matrix A, defined as $D^{-1/2}AD^{−1/2}$, being the most widely adopted one, where D is the degree matrix. However, some empirical studies show that row normalization $D^{−1}A$ outperforms it in node classification. Despite the widespread use of GNNs, there is no rigorous theoretical study on the representation power of these convolution operators, that could explain this behavior. In this work, we analyze the influence of the graph convolutions theoretically using Graph Neural Tangent Kernel in a semi-supervised node classification setting. Under a Degree Corrected Stochastic Block Model, we prove that: (i) row normalization preserves the underlying class structure better than other graph convolutions; (ii) performance degrade s with network depth due to over-smoothing, but the loss in class information is the slowest in row normalization; (iii) skip connections retain the class information even at infinite depth, thereby eliminating over-smoothing. We finally validate our theoretical findings on real datasets.

****

Bidirectional Language Models Are Also Few-shot Learners

Ajay Patel,Bryan Li,Mohammad Sadegh Rasooli,Noah Constant,Colin Raffel,Chris Callison-Burch

Large language models such as GPT-3 (Brown et al., 2020) can perform arbitrary tasks without undergoing fine-tuning after being prompted with only a few labeled examples. An arbitrary task can be reformulated as a natural language prompt, and a language model can be asked to generate the completion, indirectly performing the task in a paradigm known as prompt-based learning. To date, emergent prompt-based learning capabilities have mainly been demonstrated for unidirectional language models. However, bidirectional language models pre-trained on denoising objectives such as masked language modeling produce stronger learned representations for transfer learning. This motivates the possibility of prompting bidirectional models, but their pre-training objectives have made them largely incompatible with the existing prompting paradigm. We present SAP (Sequential Autoregressive Prompting), a technique that enables the prompting of bidirectional models. Utilizing the machine translation task as a case study, we prompt the bidirectional mT5 model (Xue et al., 2021) with SAP and demonstrate its few-shot and zero-shot translations outperform the few-shot translations of unidirectional models like GPT-3 and XGLM (Lin et al., 2021), despite mT5's approximately 50% fewer parameters. We further show SAP is effective on question answering and summarization. For the first time, our results demonstrate prompt-based learning is an emergent property of a broader class of language models, rather than only unidirectional models.

****

Revisiting adapters with adversarial training

Sylvestre-Alvise Rebuffi,Francesco Croce,Sven Gowal

While adversarial training is generally used as a defense mechanism, recent works show that it can also act as a regularizer. By co-training a neural network on clean and adversarial inputs, it is possible to improve classification accuracy on the clean, non-adversarial inputs. We demonstrate that, contrary to previous findings, it is not necessary to separate batch statistics when co-training on clean and adversarial inputs, and that it is sufficient to use adapters with few domain-specific parameters for each type of input. We establish that using the classification token of a Vision Transformer (ViT) as an adapter is enough to match the classification performance of dual normalization layers, while using significantly less additional parameters. First, we improve upon the top-1 accuracy of a non-adversarially trained ViT-B16 model by +1.12% on ImageNet (reaching 83.76% top-1 accuracy). Second, and more importantly, we show that training with adapters enables model soups through linear combinations of the clean and adversarial tokens. These model soups, which we call adversarial model soups, allow us to trade-off between clean and robust accuracy without sacrificing efficiency. Finally, we show that we can easily adapt the resulting models in the face of distribution shifts. Our ViT-B16 obtains top-1 accuracies on ImageNet variants that are on average +4.00% better than those obtained with Masked Autoencoders.

****

Human-AI Coordination via Human-Regularized Search and Learning

Hengyuan Hu,David J Wu,Adam Lerer,Jakob Nicolaus Foerster,Noam Brown

We consider the problem of making AI agents that collaborate well with humans in partially observable fully cooperative environments given datasets of human behavior. Inspired by piKL, a human-data-regularized search method that improves upon a behavioral cloning policy without diverging far away from it, we develop a three-step algorithm that achieve strong performance in coordinating with real humans in the Hanabi benchmark. We first use a regularized search algorithm and behavioral cloning to produce a better human model that captures diverse skill levels. Then, we integrate the policy regularization idea into reinforcement learn ing to train a human-like best response to the human model. Finally, we apply regularized search on top of the best response policy at test time to handle out-of-distribution challenges when playing with humans. We evaluate our method in two large scale experiments with humans. First, we show that our method outperforms experts when playing with a group of diverse human players in ad-hoc teams. Second, we show that our method beats a vanilla best response to behavioral cloning baseline by having experts play repeatedly with the two agents.

****

Solving Math Word Problems with Process-based and Outcome-based Feedback

Jonathan Uesato,Nate Kushman,Ramana Kumar,H. Francis Song,Noah Yamamoto Siegel,Lisa Wang,Antonia Creswell,Geoffrey Irving,Irina Higgins

Recent work has shown that prompting language models to generate reasoning steps improves performance on many reasoning tasks. When moving beyond prompting, this raises the question of how we should supervise the finetuning of such models: outcome-based approaches which supervise the final result, or process-based approaches which supervise the reasoning process itself? Differences between these approaches might naturally be expected not just in final-answer errors but also in reasoning errors, which can be difficult to detect and are problematic in many real-world domains such as education. We run the first comprehensive comparison between process- and outcome-based approaches trained on a natural language task, GSM8K. We find that pure outcome-based supervision produces similar final-answer error rates with less label supervision. However, for correct reasoning steps we find it necessary to use process-based supervision or supervision from learned reward models that emulate process-based feedback. In total, we improve the previous best results from 16.8% $\rightarrow$ 12.7% final-answer error and 14.0% $\rightarrow$ 3.4% reasoning error among final-answer-correct solutions.

****

EPISODE: Episodic Gradient Clipping with Periodic Resampled Corrections for Federated Learning with Heterogeneous Data

Michael Crawshaw,Yajie Bao,Mingrui Liu

Gradient clipping is an important technique for deep neural networks with exploding gradients, such as recurrent neural networks. Recent studies have shown that the loss functions of these networks do not satisfy the conventional smoothness condition, but instead satisfy a relaxed smoothness condition, i.e., the Lipschitz constant of the gradient scales linearly in terms of the gradient norm. Due to this observation, several gradient clipping algorithms have been developed for nonconvex and relaxed-smooth functions. However, the existing algorithms only apply to the single-machine or multiple-machine setting with homogeneous data across machines. It remains unclear how to design provably efficient gradient clipping algorithms in the general Federated Learning (FL) setting with heterogeneous data and limited communication rounds. In this paper, we design EPISODE, the very first algorithm to solve FL problems with heterogeneous data in the nonconvex and relaxed smoothness setting. The key ingredients of the algorithm are two new techniques called \textit{episodic gradient clipping} and \textit{periodic resampled corrections}. At the beginning of each round, EPISODE resamples stochastic gradients from each client and obtains the global averaged gradient, which is used to (1) determine whether to apply gradient clipping for the entire round and (2) construct local gradient corrections for each client. Notably, our algorithm and analysis provide a unified framework for both homogeneous and heterogeneous data under any noise level of the stochastic gradient, and it achieves state-of-the-art complexity results. In particular, we prove that EPISODE can achieve linear speedup in the number of machines, and it requires significantly fewer communication rounds. Experiments on several heterogeneous datasets, including text classification and image classification, show the superior performance of EPISODE over several strong baselines in FL. The code is available at https://github.com/MingruiLiu-ML-Lab/episode.

****

Memory-Efficient Reinforcement Learning with Priority based on Surprise and On-policyness

Ryosuke Unno,Yoshimasa Tsuruoka

In off-policy reinforcement learning, an agent collects transition data (a.k.a. experience tuples) from the environment and stores them in a replay buffer for the incoming parameter updates. Storing those tuples consumes a large amount of memory when the environment observations are given as images. Large memory consumption is especially problematic when reinforcement learning methods are applied in scenarios where the computational resources are limited. In this paper, we introduce a method to prune relatively unimportant experience tuples by a simple metric that estimates the importance of experiences and saves the overall memory consumption by the buffer. To measure the importance of experiences, we use $\textit{surprise}$ and $\textit{on-policyness}$. Surprise is quantified by the information gain the model can obtain from the experiences and on-policyness ensures that they are relevant to the current policy. In our experiments, we empirically show that our method can significantly reduce the memory consumption by the replay buffer without decreasing the performance in vision-based environments.

****

Uncovering Directions of Instability via Quadratic Approximation of Deep Neural Loss in Reinforcement Learning

Ezgi Korkmaz,Jonah Brown-Cohen

Learning in MDPs with highly complex state representations is currently possible due to multiple advancements in reinforcement learning algorithm design. However, this incline in complexity, and furthermore the increase in the dimensions of the observation came at the cost of non-robustness that can be taken advantage of (i.e. moving along worst-case directions in the observation space). To solve this policy instability problem we propose a novel method to ascertain the presence of these non-robust directions via quadratic approximation of the deep neural policy loss. Our method provides a theoretical basis for the fundamental cut-off between stable observations and non-robust observations. Furthermore, our technique is computationally efficient, and does not depend on the methods used to produce the worst-case directions. We conduct extensive experiments in the Arcade Learning Environment with several different non-robust alteration techniques. Most significantly, we demonstrate the effectiveness of our approach even in the setting where alterations are explicitly optimized to circumvent our proposed method.

****

A Theory of Dynamic Benchmarks

Ali Shirali,Rediet Abebe,Moritz Hardt

Dynamic benchmarks interweave model fitting and data collection in an attempt to mitigate the limitations of static benchmarks. In contrast to an extensive theoretical and empirical study of the static setting, the dynamic counterpart lags behind due to limited empirical studies and no apparent theoretical foundation to date. Responding to this deficit, we initiate a theoretical study of dynamic benchmarking. We examine two realizations, one capturing current practice and the other modeling more complex settings. In the first model, where data collection and model fitting alternate sequentially, we prove that model performance improves initially but can stall after only three rounds. Label noise arising from, for instance, annotator disagreement leads to even stronger negative results. Our second model generalizes the first to the case where data collection and model fitting have a hierarchical dependency structure. We show that this design guarantees strictly more progress than the first, albeit at a significant increase in complexity. We support our theoretical analysis by simulating dynamic benchmarks on two popular datasets. These results illuminate the benefits and practical limitations of dynamic benchmarking, providing both a theoretical foundation and a causal explanation for observed bottlenecks in empirical work.

****

On the Trade-Off between Actionable Explanations and the Right to be Forgotten

Martin Pawelczyk,Tobias Leemann,Asia Biega,Gjergji Kasneci

As machine learning (ML) models are increasingly being deployed in high-stakes applications, policymakers have suggested tighter data protection regulations (e.g., GDPR, CCPA). One key principle is the “right to be forgotten” which gives users the right to have their data deleted. Another key principle is the right to an actionable explanation, also known as algorithmic recourse, allowing users to reverse unfavorable decisions. To date, it is unknown whether these two principles can be operationalized simultaneously. Therefore, we introduce and study the problem of recourse invalidation in the context of data deletion requests. More specifically, we theoretically and empirically analyze the behavior of popular state-of-the-art algorithms and demonstrate that the recourses generated by these algorithms are likely to be invalidated if a small number of data deletion requests (e.g., 1 or 2) warrant updates of the predictive model. For the setting of differentiable models, we suggest a framework to identify a minimal subset of critical training points which, when removed, maximize the fraction of invalidated recourses.Using our framework, we empirically show that the removal of as little as 2 data instances from the training set can invalidate up to 95 percent of all recourses output by popular state-of-the-art algorithms. Thus, our work raises fundamental questions about the compatibility of ``the right to an actionable explanation'' in the context of the ``right to be forgotten'', while also providing constructive insights on the determining factors of recourse robustness.

****

Learning to Cooperate and Communicate Over Imperfect Channels

Jannis Weil,Gizem Ekinci,Heinz Koeppl,Tobias Meuser

Information exchange in multi-agent systems improves the cooperation among agents, especially in partially observable settings. This can be seen as part of the problem in which the agents learn how to communicate and to solve a shared task simultaneously. In the real world, communication is often carried out over imperfect channels and this requires the agents to deal with uncertainty due to potential information loss. In this paper, we consider a cooperative multi-agent system where the agents act and exchange information in a decentralized manner using a limited and unreliable channel. To cope with such channel constraints, we propose a novel communication approach based on independent Q-learning. Our method allows agents to dynamically adapt how much information to share by sending messages of different size, depending on their local observations and the channel properties. In addition to this message size selection, agents learn to encode and decode messages to improve their policies. We show that our approach outperforms approaches without adaptive capabilities and discuss its limitations in different environments.

****

Uncertainty-aware off policy learning

Xiaoying Zhang,Junpu Chen,Hongning Wang,Hong Xie,Hang Li

Off-policy learning, referring to the procedure of policy optimization with access only to logged feedback data, has shown importance in various real-world applications, such as search engines, recommender systems, etc. While the ground-truth logging policy, which generates the logged data, is usually unknown, previous work directly takes its estimated value in off-policy learning, resulting in a biased estimator. This estimator has both high bias and variance on samples with small and inaccurate estimated logging probabilities.

In this work, we explicitly model the uncertainty in the estimated logging policy and propose a novel \underline{U}ncertainty-aware \underline{I}nverse \underline{P}ropensity \underline{S}core estimator (UIPS) for improved off-policy learning. Experiment results on synthetic and three real-world recommendation datasets demonstrate the advantageous sample efficiency of the proposed UIPS estimator.

****

Renamer: A Transformer Architecture In-variant to Variable Renaming

Zachary Ankner,Alex Renda,Michael Carbin

Modeling tasks often take inputs from languages including programming languages and natural language. Many such tasks involve learning functions which are invariant to certain types of input transformations. In this work we consider a specific class of invariance: semantics-preserving variable renaming. We first show that transformer networks trained on such tasks do not always mirror the invariance of the underlying function. In this work we propose Renamer, a transformer architecture which is invariant to semantics-preserving variable renaming. Renamer improves over a vanilla transformer by between a 24.79% to 52.80% reduction in error on a case study on learning a surrogate of a large-scale CPU simualtor. Furthermore, the invariant network does not experience the same sensitivity to variable renaming, and its error remains constant when evaluated on a variable renamed version of the test set. Finally, the invariant network is more efficient to train, and matches the best error of the vanilla network with a between 25.15% to 60.00% reduction in training epochs.

****

Learning What and Where: Disentangling Location and Identity Tracking Without Supervision

Manuel Traub,Sebastian Otte,Tobias Menge,Matthias Karlbauer,Jannik Thuemmel,Martin V. Butz

Our brain can almost effortlessly decompose visual data streams into background and salient objects. Moreover, it can anticipate object motion and interactions, which are crucial abilities for conceptual planning and reasoning. Recent object reasoning datasets, such as CATER, have revealed fundamental shortcomings of current vision-based AI systems, particularly when targeting explicit object representations, object permanence, and object reasoning. Here we introduce a self-supervised LOCation and Identity tracking system (Loci), which excels on the CATER tracking challenge. Inspired by the dorsal and ventral pathways in the brain, Loci tackles the binding problem by processing separate, slot-wise encodings of 'what' and 'where'. Loci's predictive coding-like processing encourages active error minimization, such that individual slots tend to encode individual objects. Interactions between objects and object dynamics are processed in the disentangled latent space. Truncated backpropagation through time combined with forward eligibility accumulation significantly speeds up learning and improves memory efficiency. Besides exhibiting superior performance in current benchmarks, Loci effectively extracts objects from video streams and separates them into location and Gestalt components. We believe that this separation offers a representation that will facilitate effective planning and reasoning on conceptual levels.

****

BALTO: fast tensor program optimization with diversity-based active learning

Jun Bi,Xiaqing Li,Qi Guo,Rui Zhang,Yuanbo Wen,Xing Hu,Zidong Du,Xinkai Song,Yifan Hao,Yunji Chen

Tensor program optimization (TPO) based on pre-trained models can effectively reduce the computing time of deep neural networks. However, training of such models is prohibitively expensive, which highly depends on a large-scale dataset and thus requires tremendous time-consuming performance measurements (more than 1 million) on target platforms. In this paper, we propose BALTO, a fast TPO approach with biased-diversity-based active learning, aiming at reducing much lower training costs under similar optimization accuracy.The key insight is that random sampling of existing approaches suffers from a heavy redundancy of low-performance programs, which incurs tremendous duplicated time-consuming measurements. Inspired by this, BALTO removes such redundancy by introducing active learning (AL) to TPO for a much lower training cost. However, applying AL with a brute-force way in BALTO can lead to an overestimation problem. To address this, we further propose a biased-diversity-based diversity scheme specially designed for BALTO. We compare BALTO against TenSet on 6 typical hardware platforms over 2 learning models. Experimental results show that, on average, BALTO only requires 5% of the total performance measurements of TenSet to achieve the same or higher model accuracy. Moreover, the optimized tensor programs even outperform that of TenSet by 1.06% due to higher model accuracy.

****

RoCourseNet: Distributionally Robust Training of a Prediction Aware Recourse Model

Hangzhi Guo,Feiran Jia,Jinghui Chen,Anna Squicciarini,Amulya Yadav

Counterfactual (CF) explanations for machine learning (ML) models are preferred by end-users, as they explain the predictions of ML models by providing a recourse (or contrastive) case to individuals who are adversely impacted by predicted outcomes. Existing CF explanation methods generate recourses under the assumptio n that the underlying target ML model remains stationary over time. However, due to commonly occurring distributional shifts in training data, ML models constantly get updated in practice, which might render previously generated recourses invalid and diminish end-users trust in our algorithmic framework. To address this problem, we propose RoCourseNet, a training framework that jointly optimizes for predictions and recourses that are robust to future data shifts. We have three main contributions: (i) We propose a novel \emph{virtual data shift (VDS)} algorithm to find worst-case shifted ML models by explicitly considering the worst-case data shift in the training dataset. (ii) We leverage adversarial training to solve a novel tri-level optimization problem inside RoCourseNet, which simultaneously generates predictions and corresponding robust recourses. (iii) Finally, we evaluate RoCourseNet's performance on three real-world datasets and show that RoCourseNet outperforms state-of-the-art baselines by $\sim$10\% in generating robust CF explanations.

****

Inducing Meaningful Units from Character Sequences with Dynamic Capacity Slot Attention

Melika Behjati,James Henderson

Characters do not convey meaning, but sequences of characters do. We propose an unsupervised distributional method to learn the abstract meaning-bearing units in a sequence of characters. Rather than segmenting the sequence, our Dynamic Capacity Slot Attention model discovers continuous representations of the \textit{objects} in the sequence, extending an architecture for object discovery in images. We train our model on different languages and evaluate the quality of the obtained representations with forward and reverse probing classifiers. These experiments show that our model succeeds in discovering units which are similar to those proposed previously in form, content and level of abstraction, and which show promise for capturing meaningful information at a higher level of abstraction.

****

In-context Reinforcement Learning with Algorithm Distillation

Michael Laskin,Luyu Wang,Junhyuk Oh,Emilio Parisotto,Stephen Spencer,Richie Steigerwald,DJ Strouse,Steven Stenberg Hansen,Angelos Filos,Ethan Brooks,maxime gazeau,Himanshu Sahni,Satinder Singh,Volodymyr Mnih

We propose Algorithm Distillation (AD), a method for distilling reinforcement learning (RL) algorithms into neural networks by modeling their training histories with a causal sequence model. Algorithm Distillation treats learning to reinforcement learn as an across-episode sequential prediction problem. A dataset of learning histories is generated by a source RL algorithm, and then a causal transformer is trained by autoregressively predicting actions given their preceding learning histories as context. Unlike sequential policy prediction architectures that distill post-learning or expert sequences, AD is able to improve its policy entirely in-context without updating its network parameters. We demonstrate that AD can reinforcement learn in-context in a variety of environments with sparse rewards, combinatorial task structure, and pixel-based observations, and find that AD learns a more data-efficient RL algorithm than the one that generated the source data.

****

Computing all Optimal Partial Transports

Abhijeet Phatak,Sharath Raghvendra,CHITTARANJAN TRIPATHY,Kaiyi Zhang

We consider the classical version of the optimal partial transport problem. Let

μ (with a mass of U) and ν (with a mass of S) be two discrete mass distributions with $S \le U$ and let n be the total number of points in the supports of μ and ν. For a parameter $\alpha \in [0,S]$, consider the minimum-cost transport plan σ_α that transports a mass of α from ν to μ. An \emph{OT-profile} captures the behavior of the cost of σ_α as α varies from 0 to S. There is only limited work on OT-profile and its mathematical properties (see~\cite{figalli2010optimal}). In this paper, we present a novel framework to analyze the properties of the OT-profile and also present an algorithm to compute it. When μ and ν are discrete mass distributions, we show that the OT-profile is a piecewise-linear non-decreasing convex function. Let K be the combinatorial complexity of this function, i.e., the number of line segments required to represent the OT-profile. Our exact algorithm computes the OT-profile in $\tilde{O}(n^2K)$ time. Given $\delta > 0$, we also show that the algorithm by ~\cite{lahn2019graph} can be used to δ-approximate the OT-profile in $O(n^2/\delta + n/\delta^2)$ time. This approximation is a piecewise-linear function of a combinatorial complexity of $O(1/\delta)$.

An OT-profile is arguably more valuable than the OT-cost itself and can be used within applications. Under a reasonable assumption of outliers, we also show that the first derivative of the OT-profile sees a noticeable rise before any of the mass from outliers is transported. By using this property, we get an improved prediction accuracy for an outlier detection experiment. We also use this property to predict labels and estimate the class priors within PU-Learning experiments. Both these experiments are conducted on real datasets.

****

Towards Federated Learning of Deep Graph Neural Networks

Zhihua Tian,Yuan Ding,Rui Zhang,Jian Liu,Kui Ren

Graph neural networks (GNNs) learn node representations by recursively aggregating neighborhood information on graph data. However, in the federated setting, data samples (nodes) located in different clients may be connected to each other, leading to huge information loss to the training method.

Existing federated graph learning frameworks solve such a problem by generating missing neighbors or sending information across clients directly. None are suitable for training deep GNNs, which require a more expansive receptive field and higher communication costs.

In this work, we introduce a novel framework named Fed^2GNN for federated graph learning of deep GNNs via reconstructing neighborhood information of nodes. Specifically, we design a graph structure named rooted tree. The node embedding obtained by encoding on the rooted tree is the same as that obtained by encoding on the induced subgraph surrounding the node, which allows us to reconstruct the neighborhood information by building the rooted tree of the node. An encoder-decoder framework is then proposed, wherein we first encode missing neighbor information and then decode it to build the rooted tree.

Extensive experiments on real-world network datasets show the effectiveness of our framework for training deep GNNs while also achieving better performance for training shadow GNN models

****

CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations

Hangzhi Guo,Thanh Hong Nguyen,Amulya Yadav

Counterfactual (or CF) explanations are a type of local explanations for Machine Learning (ML) model predictions, which offer a contrastive case as an explanation by finding the smallest changes (in feature space) to the input data point, which will lead to a different prediction by the ML model. Existing CF explanation techniques suffer from two major limitations: (i) all of them are post-hoc methods designed for use with proprietary ML models --- as a result, their procedure for generating CF explanations is uninformed by the training of the ML model, which leads to misalignment between model predictions and explanations; and (ii) most of them rely on solving separate time-intensive optimization problems to find CF explanations for each input data point (which negatively impacts their runtime). This work makes a novel departure from the prevalent post-hoc paradigm (

of generating CF explanations) by presenting CounterNet, an end-to-end learning framework which integrates predictive model training and the generation of counterfactual (CF) explanations into a single pipeline. We adopt a block-wise coordinate descent procedure which helps in effectively training CounterNet's network. Our extensive experiments on multiple real-world datasets show that CounterNet generates high-quality predictions, and consistently achieves 100% CF validity and very low proximity scores (thereby achieving a well-balanced cost-invalidity trade-off) for any new input instance, and runs 3X faster than existing state-of-the-art baselines.

****

SmilesFormer: Language Model for Molecular Design

Joshua Owoyemi,Nazim Medzhidov

The objective of drug discovery is to find novel compounds with desirable chemical properties. Generative models have been utilized to sample molecules at the intersection of multiple property constraints. In this paper we pose molecular design as a language modeling problem where the model implicitly learns the vocabulary and composition of valid molecules, hence it is able to generate new molecules of interest. We present SmilesFormer, a Transformer-based model which is able to encode molecules, molecule fragments, and fragment compositions as latent variables, which are in turn decoded to stochastically generate novel molecules. This is achieved by fragmenting the molecules into smaller combinatorial groups, then learning the mapping between the input fragments and valid SMILES sequences.

The model is able to optimize molecular properties through a stochastic latent space traversal technique. This technique systematically searches the encoded latent space to find latent vectors that are able to produce molecules to meet the multi-property objective. The model was validated through various de novo molecular design tasks, achieving state-of-the-art performances when compared to previous methods. Furthermore, we used the proposed method to demonstrate a drug rediscovery pipeline for Donepezil, a known Acetylcholinesterase Inhibitor.

****

AE-FLOW: Autoencoders with Normalizing Flows for Medical Images Anomaly Detection

Yuzhong Zhao,Qiaoqiao Ding,Xiaoqun Zhang

Anomaly detection from medical images is an important task for clinical screening and diagnosis. In general, a large dataset of normal images are available while only few abnormal images can be collected in clinical practice. By mimicking the diagnosis process of radiologists, we attempt to tackle this problem by learning a tractable distribution of normal images and identify anomalies by differentiating the original image and the reconstructed normal image. More specifically, we propose a normalizing flow-based autoencoder for an efficient and tractable representation of normal medical images. The anomaly score consists of the likelihood originated from the normalizing flow and the reconstruction error of the autoencoder, which allows to identify the abnormality and provide an interpretability at both image and pixel levels. Experimental evaluation on two medical images datasets showed that the proposed model outperformed the other approaches by a large margin, which validated the effectiveness and robustness of the proposed method.

****

Learning a Domain-Agnostic Policy through Adversarial Representation Matching for Cross-Domain Policy Transfer

Hayato Watahiki,Ryo Iwase,Ryosuke Unno,Yoshimasa Tsuruoka

The low transferability of learned policies is one of the most critical problems limiting the applicability of learning-based solutions to decision-making tasks. In this paper, we present a way to align latent representations of states and actions between different domains by optimizing an adversarial objective. We train two models, a policy and a domain discriminator, with unpaired trajectories of proxy tasks through behavioral cloning as well as adversarial training. After the latent representations are aligned between domains, a domain-agnostic part o

f the policy trained with any method in the source domain can be immediately transferred to the target domain in a zero-shot manner. We empirically show that our simple approach achieves comparable performance to the latest methods in zero-shot cross-domain transfer. We also observe that our method performs better than other approaches in transfer between domains with different complexities, whereas other methods fail catastrophically.
****

A Self-Attention Ansatz for Ab-initio Quantum Chemistry
Ingrid von Glehn,James S Spencer,David Pfau

We present a novel neural network architecture using self-attention, the Wavefunction Transformer (PsiFormer), which can be used as an approximation (or "Ansatz") for solving the many-electron Schrödinger equation, the fundamental equation for quantum chemistry and material science. This equation can be solved *from first principles*, requiring no external training data. In recent years, deep neural networks like the FermiNet and PauliNet have been used to significantly improve the accuracy of these first-principle calculations, but they lack an attention-like mechanism for gating interactions between electrons. Here we show that the PsiFormer can be used as a drop-in replacement for these other neural networks, often dramatically improving the accuracy of the calculations. On larger molecules especially, the ground state energy can be improved by dozens of kcal/mol, a qualitative leap over previous methods. This demonstrates that self-attention networks can learn complex quantum mechanical correlations between electrons, and are a promising route to reaching unprecedented accuracy in chemical calculations on larger systems.
****

Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse
Martin Pawelczyk,Teresa Datta,Johan Van den Heuvel,Gjergji Kasneci,Himabindu Lakkaraju

As machine learning models are increasingly being employed to make consequential decisions in real-world settings, it becomes critical to ensure that individuals who are adversely impacted (e.g., loan denied) by the predictions of these models are provided with a means for recourse. While several approaches have been proposed to construct recourses for affected individuals, the recourses output by these methods either achieve low costs (i.e., ease-of-implementation) or robustness to small perturbations (i.e., noisy implementations of recourses), but not both due to the inherent trade-offs between the recourse costs and robustness. Furthermore, prior approaches do not provide end users with any agency over navigating the aforementioned trade-offs. In this work, we address the above challenges by proposing the first algorithmic framework which enables users to effectively manage the recourse cost vs. robustness trade-offs. More specifically, our framework Probabilistically ROBust rEcourse (PROBE) lets users choose the probability with which a recourse could get invalidated (recourse invalidation rate) if small changes are made to the recourse i.e., the recourse is implemented somewhat noisily. To this end, we propose a novel objective function which simultaneously minimizes the gap between the achieved (resulting) and desired recourse invalidation rates, minimizes recourse costs, and also ensures that the resulting recourse achieves a positive model prediction. We develop novel theoretical results to characterize the recourse invalidation rates corresponding to any given instance w.r.t. different classes of underlying models (e.g., linear models, tree based models etc.), and leverage these results to efficiently optimize the proposed objective. Experimental evaluation with multiple real world datasets demonstrates the efficacy of the proposed framework.
****

How robust is unsupervised representation learning to distribution shift?
Yuge Shi,Imant Daunhawer,Julia E Vogt,Philip Torr,Amartya Sanyal

The robustness of machine learning algorithms to distributions shift is primarily discussed in the context of supervised learning (SL). As such, there is a lack of insight on the robustness of the representations learned from unsupervised methods, such as self-supervised learning (SSL) and auto-encoder based algorithms

(AE), to distribution shift. We posit that the input-driven objectives of unsupervised algorithms lead to representations that are more robust to distribution shift than the target-driven objective of SL. We verify this by extensively evaluating the performance of SSL and AE on both synthetic and realistic distribution shift datasets. Following observations that the linear layer used for classification itself can be susceptible to spurious correlations, we evaluate the representations using a linear head trained on a small amount of out-of-distribution (OOD) data, to isolate the robustness of the learned representations from that of the linear head. We also develop “controllable” versions of existing realistic domain generalisation datasets with adjustable degrees of distribution shifts. This allows us to study the robustness of different learning algorithms under versatile yet realistic distribution shift conditions. Our experiments show that representations learned from unsupervised learning algorithms generalise better than SL under a wide variety of extreme as well as realistic distribution shifts.

****

Autoregressive Generative Modeling with Noise Conditional Maximum Likelihood Estimation

Henry Li,Yuval Kluger

We introduce a simple modification to the standard maximum likelihood estimation (MLE) framework. Rather than maximizing a single unconditional likelihood of the data under the model, we maximize a family of \textit{noise conditional} likelihoods consisting of the data perturbed by a continuum of noise levels. We find that models trained this way are more robust to noise, obtain higher test likelihoods, and generate higher quality images. They can also be sampled from via a novel score-based sampling scheme which combats the classical \textit{covariate shift} problem that occurs during sample generation in autoregressive models. Applying this augmentation to autoregressive image models, we obtain 3.32 bits per dimension on the ImageNet 64x64 dataset, and substantially improve the quality of generated samples in terms of the Frechet Inception distance (FID) --- from 37.50 to 12.09 on the CIFAR-10 dataset.

****

Multi-Behavior Dynamic Contrastive Learning for Recommendation

Wei Wei,Chao Huang,Lianghao Xia,Yanwei Yu,Chuxu Zhang

Dynamic behavior modeling has become an essential task in personalized recommender systems for learning the time-evolving user preference in online platforms. However, most next-item recommendation methods follow the single type behavior learning manner, which notably limits their user representation performance in reality, since the user-item relationships are often multi-typed in real-life applications (e.g., click, tag-as-favorite, review and purchase). To offer better recommendations, this work proposes Evolving Graph Contrastive Memory Network (EGCM) to model dynamic interaction heterogeneity for multi-behavior sequential recommendation. Specifically, we first develop a multi-behavior graph encoder to capture the short-term preference heterogeneity, and preserve the dedicated relation semantics for different types of user-item interactions. In addition, we design a dynamic cross-relational memory network, empowering EGCM to distill the long-term multi-behavior preference of users and the underlying evolving cross-type behavior dependencies over time. To enhance the user representation with multi-behavior commonality and diversity, we design a multi-behavior contrastive learning paradigm with heterogeneous short- and long-term interest modeling. Experiments on several real-world datasets show the superiority of our recommender system over various state-of-the-art baselines.

****

Analyzing diffusion as serial reproduction

Raja Marjieh,Ilia Sucholutsky,Thomas A Langlois,Nori Jacoby,Thomas L. Griffiths

Diffusion models are a class of generative models that learn to synthesize samples by inverting a diffusion process that gradually maps data into noise. While these models have enjoyed great success recently, a full theoretical understanding of their observed properties is still lacking, in particular, their weak sensi tivity to the choice of noise family and the role of adequate scheduling of noise levels for good synthesis. By identifying a correspondence between diffusion models and a well-known paradigm in cognitive science known as serial reproduction, whereby human agents iteratively observe and reproduce stimuli from memory, we show how the aforementioned properties of diffusion models can be explained as a natural consequence of this correspondence. We then complement our theoretical analysis with simulations that exhibit these key features. Our work highlights how classic paradigms in cognitive science can shed light on state-of-the-art machine learning problems.

****

Pseudo-label Training and Model Inertia in Neural Machine Translation

Benjamin Hsu,Anna Currey,Xing Niu,Maria Nadejde,Georgiana Dinu

Like many other machine learning applications, neural machine translation (NMT) benefits from over-parameterized deep neural models. However, these models have been observed to be brittle: NMT model predictions are sensitive to small input changes and can show significant variation across re-training or incremental model updates. This work studies a frequently used method in NMT, pseudo-label training (PLT), which is common to the related techniques of forward-translation (or self-training) and sequence-level knowledge distillation. While the effect of PLT on quality is well-documented, we highlight a lesser-known effect: PLT can enhance a model's stability to model updates and input perturbations, a set of properties we call \textit{model inertia}. We study inertia effects under different training settings and we identify distribution simplification as a mechanism behind the observed results.

****

Adaptive Smoothing Gradient Learning for Spiking Neural Networks

Ziming Wang,Runhao Jiang,Shuang Lian,Rui Yan,Huajin Tang

Spiking neural networks (SNNs) with biologically inspired spatio-temporal dynamics show higher energy efficiency on neuromorphic architectures. Error backpropagation in SNNs is prohibited by the all-or-none nature of spikes. The existing solution circumvents this problem by a relaxation on the gradient calculation using a continuous function with a constant relaxation degree, so-called surrogate gradient learning. Nevertheless, such solution introduces additional smoothness error on spiking firing which leads to the gradients being estimated inaccurately. Thus, how to adjust the relaxation degree adaptively and eliminate smoothness error progressively is crucial. Here, we propose a methodology such that training a prototype neural network will evolve into training an SNN gradually by fusing the learnable relaxation degree into the network with random spike noise. In this way, the network learns adaptively the accurate gradients of loss landscape in SNNs. The theoretical analysis further shows optimization on such a noisy network could be evolved into optimization on the embedded SNN with shared weights progressively. Moreover, we conduct extensive experiments on static images, dynamic event streams, speech, and instrumental sounds. The results show the proposed method achieves state-of-the-art performance across all the datasets with remarkable robustness on different relaxation degrees.

****

Going Beyond Approximation: Encoding Constraints for Explainable Multi-hop Inference via Differentiable Combinatorial Solvers

Mokanarangan Thayaparan,Marco Valentino,Andre Freitas

Integer Linear Programming (ILP) provides a viable mechanism to encode explicit and controllable assumptions about explainable multi-hop inference with natural language. However, an ILP formulation is non-differentiable and cannot be integrated into broader deep learning architectures. Recently, Thayaparan et al. (2021a) proposed a novel methodology to integrate ILP with Transformers to achieve end-to-end differentiability for complex multi-hop inference. While this hybrid framework has been demonstrated to deliver better answer and explanation selection than transformer-based and existing ILP solvers, the neuro-symbolic integration still relies on a convex relaxation of the ILP formulation, which can produce sub-optimal solutions. To improve these limitations, we propose Diff-Comb Explainer, a novel neuro-symbolic architecture based on Differentiable BlackBox Combina torial solvers (DBCS) (Pogan■i■ et al., 2019). Unlike existing differentiable solvers, the presented model does not require the transformation and relaxation of the explicit semantic constraints, allowing for direct and more efficient integration of ILP formulations. Diff-Comb Explainer demonstrates improved accuracy and explainability over non-differentiable solvers, Transformers and existing differentiable constraint-based multi-hop inference frameworks.

****

Robust and accelerated single-spike spiking neural network training with applicability to challenging temporal tasks

Luke Taylor,Andrew J King,Nicol Spencer Harper

Spiking neural networks (SNNs), particularly the single-spike variant in which neurons spike at most once, are considerably more energy efficient than standard artificial neural networks (ANNs). However, single-spike SSNs are difficult to train due to their dynamic and non-differentiable nature, where current solutions are either slow or suffer from training instabilities. These networks have also been critiqued for their limited computational applicability such as being unsuitable for time-series datasets. We propose a new model for training single-spike SNNs which mitigates the aforementioned training issues and obtains competitive results across various image and neuromorphic datasets, with up to a $13.98\times$ training speedup and up to an 81% reduction in spikes compared to the multi-spike SNN. Notably, our model performs on par with multi-spike SNNs in challenging tasks involving neuromorphic time-series datasets, demonstrating a broader computational role for single-spike SNNs than previously believed.

****

A NEW PARADIGM FOR CROSS-MODALITY PERSON RE-IDENTIFICATION

Yumeng Wang,Feng Yang,Tongkai Xu,Yanze Zhu

Visible and infrared Person Re-identification(ReID) is still very challenging on account of few cross-modality dataset and large inter-modality variation. Most existing cross-modality ReID methods have trouble eliminating cross-modality discrepancy resulting from the heterogeneous images. In this paper, we present an effective framework and build a large benchmark, named NPU-ReID. To this end, we propose a dual-path fusion network and taking transformer as the smallest feature extraction unit. To expand cross-modality sample diversity, we propose a modality augmentation strategy to generate semi-modality pedestrian images by exchanging certain patch and the main innovation is that the cross-modality gap can be indirectly minimized by reducing the variance of semi-modality and infrared or visible modality. Moreover, in order to make the traditional triplet loss more suitable for cross-modal matching tasks, multi-masking triplet loss is a targeted design for optimizing the relative distance between anchor and positive/negative samples pairs from cross-modality, especially constraining the distance between simple and hard positive samples. Experimental results demonstrate that our proposed method achieves superior performance than other methods on SYSU-MM01, RegDB and our proposed NPU-ReID dataset, especially on the RegDB dataset with significant improvement of 6.81$\%$ in rank1 and 9.65$\%$ in mAP.

****

Causal Mean Field Multi-Agent Reinforcement Learning

Hao Ma,Zhiqiang Pu,Yi Pan,Boyin Liu,Min Chen,Shijie Wang

Scalability remains a challenge in multi-agent reinforcement learning and is currently under active research. However, existing works lack the ability to identify the essential interaction under the non-stationary environment. We propose causal mean field Q-learning (CMFQ) to address this problem. It has the advantage of MFQ, which can compress the space size dramatically. Besides, it is ever more robust toward the non-stationary caused by increasing agents. We enable agents to identify which ally or opponent is more crucial by asking "what if" with the help of the structural causal model (SCM), then pay more attention to more crucial ones. We test CMFQ in mixed cooperative-competitive and cooperative games, which verify our method's scalability performance.

****

Hidden Markov Mixture of Gaussian Process Functional Regression: Utilizing Multi-Scale Structure for Time-Series Forecasting

Tao Li,Jinwen Ma

The mixture of Gaussian process functional regressions (GPFRs) assumes that there are a batch of time-series or sample curves which are generated by independent random processes with different temporal structures. However, in the real situations, these structures are actually transferred in a random manner from a long time scale. Therefore, the assumption of independent curves is not true in practice. In order to get rid of this limitation, we propose the hidden Markov based GPFR mixture model (HM-GPFR) by describing these curves with both fine and coarse level temporal structures. Specifically, the temporal structure is described by the Gaussian process model at the fine level and hidden Markov process at the coarse level. The whole model can be regarded as a random process with state switching dynamics. To further enhance the robustness of the model, we also give a priori to the model parameters and develop Bayesian hidden Markov based GPFR mixture model (BHM-GPFR). Experimental results demonstrate that the proposed methods have both high prediction accuracy and good interpretability.

****

HyperDeepONet: learning operator with complex target function space using the limited resources via hypernetwork

Jae Yong Lee,SungWoong CHO,Hyung Ju Hwang

Fast and accurate predictions for complex physical dynamics are a big challenge across various applications. Real-time prediction on resource-constrained hardware is even more crucial in the real-world problems. The deep operator network (DeepONet) has recently been proposed as a framework for learning nonlinear mappings between function spaces. However, the DeepONet requires many parameters and has a high computational cost when learning operators, particularly those with complex (discontinuous or non-smooth) target functions. In this study, we propose HyperDeepONet, which uses the expressive power of the hypernetwork to enable learning of a complex operator with smaller set of parameters. The DeepONet and its variant models can be thought of as a method of injecting the input function information into the target function. From this perspective, these models can be viewed as a special case of HyperDeepONet. We analyze the complexity of DeepONet and conclude that HyperDeepONet needs relatively lower complexity to obtain the desired accuracy for operator learning. HyperDeepONet was successfully applied to various operator learning problems using low computational resources compared to other benchmarks.

****

Edge Guided GANs with Contrastive Learning for Semantic Image Synthesis

Hao Tang,XIAOJUAN QI,Guolei Sun,Dan Xu,Nicu Sebe,Radu Timofte,Luc Van Gool

We propose a novel \underline{e}dge guided \underline{g}enerative \underline{a}dversarial \underline{n}etwork with \underline{c}ontrastive learning (ECGAN) for the challenging semantic image synthesis task. Although considerable improvement has been achieved, the quality of synthesized images is far from satisfactory due to three largely unresolved challenges. 1) The semantic labels do not provide detailed structural information, making it difficult to synthesize local details and structures. 2) The widely adopted CNN operations such as convolution, down-sampling, and normalization usually cause spatial resolution loss and thus cannot fully preserve the original semantic information, leading to semantically inconsistent results (e.g., missing small objects). 3) Existing semantic image synthesis methods focus on modeling ``local'' semantic information from a single input semantic layout. However, they ignore ``global'' semantic information of multiple input semantic layouts, i.e., semantic cross-relations between pixels across different input layouts. To tackle 1), we propose to use edge as an intermediate representation which is further adopted to guide image generation via a proposed attention guided edge transfer module. Edge information is produced by a convolutional generator and introduces detailed structure information. To tackle 2), we design an effective module to selectively highlight class-dependent feature maps according to the original semantic layout to preserve the semantic information. To tackle 3), inspired by current methods in contrastive learning, we propose a novel contrastive learning method, which aims to enforce pixel embeddings belonging to the same semantic class to generate more similar image content than those from different classes. Doing so can capture more semantic relations by explicitly exploring the structures of labeled pixels from multiple input semantic layouts. Experiments on three challenging datasets show that our ECGAN achieves significantly better results than state-of-the-art methods.

****

Towards Reliable Link Prediction with Robust Graph Information Bottleneck

Zhanke Zhou,Jiangchao Yao,Jiaxu Liu,Xiawei Guo,LI He,Shuo Yuan,quanming yao,Liang Wang,Bo Han

Link prediction on graphs has achieved great success with the rise of deep graph learning. However, the potential robustness under the edge noise is less investigated. We reveal that the inherent edge noise that naturally perturbs both input topology and target label leads to severe performance degradation and representation collapse. Here, we propose an information-theory guided principle, Robust Graph Information Bottleneck (RGIB), to extract reliable supervision signals and avoid representation collapse. Different from the general information bottleneck, RGIB decouples and balances the mutual dependence among graph topology, edge label, and representation, building a new learning objective for robust representation. We also provide two implementations, RGIB-SSL and RGIB-REP, that benefit from different methodologies, i.e., self-supervised learning and data reparametrization, for indirect and direct data denoising, respectively. Extensive experiments on six benchmarks with various scenarios verify the effectiveness of the proposed RGIB.

****

Enforcing Delayed-Impact Fairness Guarantees

Aline Weber,Blossom Metevier,Yuriy Brun,Philip S. Thomas,Bruno Castro da Silva

Recent research has shown that seemingly fair machine learning models, when used to inform decisions that have an impact on people's lives or well-being (e.g., applications involving education, employment, and lending), can inadvertently increase social inequality in the long term. Existing fairness-aware algorithms consider static fairness constraints, such as equal opportunity or demographic parity, but enforcing constraints of this type may result in models that have a negative long-term impact on disadvantaged individuals and communities. We introduce ELF (Enforcing Long-term Fairness), the first classification algorithm that provides high-confidence fairness guarantees in terms of long-term, or delayed, impact. Importantly, ELF solves the open problem of providing such guarantees based only on historical data that includes observations of delayed impact. Prior methods, by contrast, require prior knowledge (or an estimate) of analytical models describing the relationship between a classifier's predictions and their corresponding delayed impact. We prove that ELF satisfies delayed-impact fairness constraints with high confidence and that it is guaranteed to identify a fair solution, if one exists, given sufficient data. We show empirically, using real-life data, that ELF can successfully mitigate long-term unfairness with high confidence.

****

Affinity-Aware Graph Networks

Ameya Velingker,Ali Kemal Sinop,Ira Ktena,Petar Veli■kovi■,Sreenivas Gollapudi

Graph Neural Networks (GNNs) have emerged as a powerful technique for learning on relational data. Owing to the relatively limited number of message passing steps they perform—and hence a smaller receptive field—there has been significant interest in improving their expressivity by incorporating structural aspects of the underlying graph. In this paper, we explore the use of affinity measures as features in graph neural networks, in particular measures arising from random walks, including effective resistance, hitting and commute times. We propose message passing networks based on these features and evaluate their performance on a variety of node and graph property prediction tasks.

****

Towards the Detection of Diffusion Model Deepfakes

Jonas Ricker,Simon Damm,Thorsten Holz,Asja Fischer

Diffusion models (DMs) have recently emerged as a promising method in image synt hesis. They have surpassed generative adversarial networks (GANs) in both diversity and quality, and have achieved impressive results in text-to-image modeling. However, to date, only little attention has been paid to the detection of DM-generated images, which is critical to prevent adverse impacts on our society. While prior works have shown that GAN-generated images can be reliably detected using automated methods, it is unclear whether the same methods are effective against DMs. In this work, we address this challenge and take a first look at detecting DM-generated images. We approach the problem from two different angles: First, we evaluate the performance of state-of-the-art detectors on a variety of DMs. Second, we analyze DM-generated images in the frequency domain and study different factors that influence the spectral properties of these images. Most importantly, we demonstrate that GANs and DMs produce images with different characteristics, which requires adaptation of existing classifiers to ensure reliable detection. We believe this work provides the foundation and starting point for further research to detect DM deepfakes effectively.

****

Global-Scale Species Mapping From Crowdsourced Data

Elijah Cole,Grant Van Horn,Alexander Shepard,Patrick Leary,Scott Loarie,Pietro Perona,Oisin Mac Aodha

Estimating the geographical range of a species from in situ observational data is a challenging and important geospatial prediction problem. Given a set of locations indicating where a species has been observed, the goal is to learn a model that can predict how likely it is for the species to be present at any other location. While this is a well-studied problem, traditional approaches are unable to take advantage of more recently available large-scale datasets that cover many locations and species. We propose a new approach that jointly estimates the geographical ranges of tens of thousands of different species simultaneously. We develop a series of benchmark evaluation tasks that measure different aspects of the species range and spatial representation learning problems. We show that our approach scales both in terms of amount of training data and species, where adding more data enables the models to learn better spatial representations that generalize to other species. Despite being only trained on weakly supervised crowdsourced data, our models can approach the predictions of current expert-developed gold standard models.

****

CANIFE: Crafting Canaries for Empirical Privacy Measurement in Federated Learning

Samuel Maddock,Alexandre Sablayrolles,Pierre Stock

Federated Learning (FL) is a setting for training machine learning models in distributed environments where the clients do not share their raw data but instead send model updates to a server. However, model updates can be subject to attacks and leak private information. Differential Privacy (DP) is a leading mitigation strategy which involves adding noise to clipped model updates, trading off performance for strong theoretical privacy guarantees. Previous work has shown that the threat model of DP is conservative and that the obtained guarantees may be vacuous or may overestimate information leakage in practice. In this paper, we aim to achieve a tighter measurement of the model exposure by considering a realistic threat model. We propose a novel method, CANIFE, that uses canaries - carefully crafted samples by a strong adversary to evaluate the empirical privacy of a training round. We apply this attack to vision models trained on CIFAR-10 and CelebA and to language models trained on Sent140 and Shakespeare. In particular, in realistic FL scenarios, we demonstrate that the empirical per-round epsilon obtained with CANIFE is 4 -- 5$\times$ lower than the theoretical bound.

****

Multivariate Time Series Forecasting By Graph Attention Networks With Theoretical Guarantees

Zhi Zhang,Weijian Li,Han Liu

Multivariate time series forecasting (MTSF) aims to predict future values of multiple variables based on past values of multivariate time series, and has been applied in fields including traffic flow prediction, stock price forecasting, and anomaly detection. Capturing the inter-dependencies among variables poses one significant challenge to MTSF. Several methods that model the correlations between variables with an aim to improve the test prediction accuracy have been considered in recent works, however, none of them have theoretical guarantees. In this paper, we developed a new norm-bounded graph attention network (GAT) for MTSF by upper-bounding the Frobenius norm of weights in each layer of the GAT model to achieve optimal performance.

Under optimal parameters, we theoretically show that our model can achieve a generalization error bound which is expressed as products of Frobenius norm of weight in each layer and the numbers of neighbors and attention heads, while the latter is represented as polynomial terms with the degree as the number of layers.

Empirically, we investigate the impact of different components of GAT models on the performance of MTSF.

Our experiment also verifies our theoretical findings. Empirically, we also observe that the generalization performance of our method is dependent on the number of attention heads, the number of neighbors, the scales (norms) of the weight matrices, the scale of the input features, and the number of layers.

Our method provides novel perspectives for improving the generation performance for MTSF, and our theoretical guarantees give substantial implications for designing attention-based methods for MTSF.

****

Maximal Correlation-Based Post-Nonlinear Learning for Bivariate Causal Discovery

Tianjian Zhang,Feng Yin,Zhi-Quan Luo

Bivariate causal discovery aims to determine the causal relationship between two random variables from passive observational data (as intervention is not affordable in many scientific fields), which is considered fundamental and challenging. Designing algorithms based on the post-nonlinear (PNL) model has aroused much attention for its generality. However, the state-of-the-art (SOTA) PNL-based algorithms involve highly non-convex objectives for neural network training, which are time-consuming and unable to produce meaningful solutions with finite samples. In this paper, we propose a novel method that incorporates maximal correlation into the PNL model learning (short as MC-PNL) such that the underlying nonlinearities can be accurately recovered. Owing to the benign structure of our objective function when modeling the nonlinearities with linear combinations of random Fourier features, the target optimization problem can be solved rather efficiently and rapidly via the block coordinate descent. We also compare the MC-PNL with SOTA methods on the downstream synthetic and real causal discovery tasks to show its superiority in time and accuracy. Our code is available at https://anonymous.4open.science/r/MC-PNL-E446/.

****

A View From Somewhere: Human-Centric Face Representations

Jerone Theodore Alexander Andrews,Przemyslaw Joniak,Alice Xiang

Few datasets contain self-identified demographic information, inferring demographic information risks introducing additional biases, and collecting and storing data on sensitive attributes can carry legal risks. Besides, categorical demographic labels do not necessarily capture all the relevant dimensions of human diversity. We propose to implicitly learn a set of continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. We uncover the dimensions by learning on A View From Somewhere (AVFS) dataset of 638,180 human judgments of face similarity. We demonstrate the utility of our learned embedding space for predicting face similarity judgments, collecting continuous face attribute values, attribute classification, and comparative dataset diversity auditing. Moreover, using a novel conditional framework, we show that an annotator's demographics influences the \emph{importance} they place on different attributes when judging similarity, underscoring the \emph{need} for diverse annotator groups to avoid biases. Data and code are available at \url{https://github.com/SonyAI/a_view_from_somewhere}.

****

Identifiability Results for Multimodal Contrastive Learning

Imant Daunhawer,Alice Bizeul,Emanuele Palumbo,Alexander Marx,Julia E Vogt

Contrastive learning is a cornerstone underlying recent progress in multi-view and multimodal learning, e.g., in representation learning with image/caption pairs. While its effectiveness is not yet fully understood, a line of recent work reveals that contrastive learning can invert the data generating process and recover ground truth latent factors shared between views. In this work, we present new identifiability results for multimodal contrastive learning, showing that it is possible to recover shared factors in a more general setup than the multi-view setting studied previously. Specifically, we distinguish between the multi-view setting with one generative mechanism (e.g., multiple cameras of the same type) and the multimodal setting that is characterized by distinct mechanisms (e.g., cameras and microphones). Our work generalizes previous identifiability results by redefining the generative process in terms of distinct mechanisms with modality-specific latent variables. We prove that contrastive learning can block-identify latent factors shared between modalities, even when there are nontrivial dependencies between factors. We empirically verify our identifiability results with numerical simulations and corroborate our findings on a complex multimodal dataset of image/text pairs. Zooming out, our work provides a theoretical basis for multimodal representation learning and explains in which settings multimodal contrastive learning can be effective in practice.

****

Federated Learning as Variational Inference: A Scalable Expectation Propagation Approach

Han Guo,Philip Greengard,Hongyi Wang,Andrew Gelman,Yoon Kim,Eric Xing

The canonical formulation of federated learning treats it as a distributed optimization problem where the model parameters are optimized against a global loss function that decomposes across client loss functions. A recent alternative formulation instead treats federated learning as a distributed inference problem, where the goal is to infer a global posterior from partitioned client data (Al-Shedivat et al., 2021). This paper extends the inference view and describes a variational inference formulation of federated learning where the goal is to find a global variational posterior that well-approximates the true posterior. This naturally motivates an expectation propagation approach to federated learning (FedEP), where approximations to the global posterior are iteratively refined through probabilistic message-passing between the central server and the clients. We conduct an extensive empirical study across various algorithmic considerations and describe practical strategies for scaling up expectation propagation to the modern federated setting. We apply FedEP on standard federated learning benchmarks and find that it outperforms strong baselines in terms of both convergence speed and accuracy.

****

Latent Graph Inference using Product Manifolds

Haitz Sáez de Ocáriz Borde,Anees Kazi,Federico Barbero,Pietro Lio

Graph Neural Networks usually rely on the assumption that the graph topology is available to the network as well as optimal for the downstream task. Latent graph inference allows models to dynamically learn the intrinsic graph structure of problems where the connectivity patterns of data may not be directly accessible. In this work, we generalize the discrete Differentiable Graph Module (dDGM) for latent graph learning. The original dDGM architecture used the Euclidean plane to encode latent features based on which the latent graphs were generated. By incorporating Riemannian geometry into the model and generating more complex embedding spaces, we can improve the performance of the latent graph inference system. In particular, we propose a computationally tractable approach to produce product manifolds of constant curvature model spaces that can encode latent features of varying structure. The latent representations mapped onto the inferred product manifold are used to compute richer similarity measures that are leveraged by the latent graph learning model to obtain optimized latent graphs. Moreover, the curvature of the product manifold is learned during training alongside the rest of the network parameters and based on the downstream task, rather than it being a static embedding space. Our novel approach is tested on a wide range of dat asets, and outperforms the original dDGM model.

****

UNICORN: A Unified Backdoor Trigger Inversion Framework

Zhenting Wang,Kai Mei,Juan Zhai,Shiqing Ma

The backdoor attack, where the adversary uses inputs stamped with triggers (e.g., a patch) to activate pre-planted malicious behaviors, is a severe threat to Deep Neural Network (DNN) models. Trigger inversion is an effective way of identifying backdoor models and understanding embedded adversarial behaviors. A challenge of trigger inversion is that there are many ways of constructing the trigger. Existing methods cannot generalize to various types of triggers by making certain assumptions or attack-specific constraints. The fundamental reason is that existing work does not formally define the trigger and the inversion problem. This work formally defines and analyzes the trigger and the inversion problem. Then, it proposes a unified framework to invert backdoor triggers based on the formalization of triggers and the identified inner behaviors of backdoor models from our analysis. Our prototype UNICORN is general and effective in inverting backdoor triggers in DNNs. The code can be found at https://github.com/RU-System-Software-and-Security/UNICORN.

****

DBA: Efficient Transformer with Dynamic Bilinear Low-Rank Attention

Bosheng Qin,Juncheng Li,Siliang Tang,Yueting Zhuang

Many studies have been conducted to improve the efficiency of the Transformer from quadric to linear over long sequence conditions. Among them, the low-rank-based methods aim to learn the projection matrices to compress the sequence length, thus achieving efficiency gain. However, the projection matrices are fixed once they have been learned, which compress the sequence length with dedicated coefficients for the tokens in the same position regardless of different sequences. Adopting such input-invariant low-rank projections ignores the fact that the most informative part of a sequence varies from sequence to sequence, thus failing to preserve the most useful information that lies in varied positions of different sequences. In addition, previous efficient Transformers only focus on the influence of sequence length while neglecting the effect of hidden state dimension to achieve further efficiency gain. To address the aforementioned problems, we present an efficient yet effective attention mechanism, namely the Dynamic Bilinear Low-Rank Attention (DBA), which compresses the sequence length by input-sensitive dynamic projection matrices and achieves linear time and space complexity by jointly optimizing the sequence length and hidden state dimension while maintaining state-of-the-art performance. Specifically, we first theoretically demonstrate that the sequence length can be compressed non-destructively from a novel perspective of the information theory, with the compression matrices dynamically determined by the input sequence. Furthermore, we show that the hidden state dimension can be approximated by extending the Johnson–Lindenstrauss lemma and achieves high-order small amount error, optimizing the attention in bilinear form. In addition, theoretical analysis shows that the DBA is proficient in capturing high-order relations in cross-attention problems. Experiments over tasks with diverse sequence length conditions show that the DBA achieves state-of-the-art performance compared with various strong baselines while maintaining less memory consumption with higher speed, demonstrating the effectiveness and efficiency of the DBA.

****

On the Robustness of Dataset Inference

Sebastian Szyller,Rui Zhang,Jian Liu,N Asokan

Machine learning (ML) models are costly to train as they can require a significant amount of data, computational resources and technical expertise. Thus, they constitute valuable intellectual property that needs protection from adversaries wanting to steal them. $\textit{Ownership verification}$ techniques allow the victims of model stealing attacks to demonstrate that a suspect model was in fact stolen from theirs.

Although a number of ownership verification techniques based on watermarking or fingerprinting have been proposed, most of them fall short either in terms of se curity guarantees (well-equipped adversaries can evade verification) or computational cost. A fingerprinting technique introduced at ICLR '21, $\textit{Dataset Inference}$ (DI), has been shown to offer better robustness and efficiency than prior methods.

The authors of DI provided a correctness proof for linear (suspect) models. However, in the same setting, we prove that DI suffers from high false positives (FPs) -- it can incorrectly identify an independent model trained with non-overlapping data from the same distribution as stolen. We further prove that DI also triggers FPs in realistic, non-linear suspect models. We then confirm empirically that DI leads to FPs, with high confidence.

Second, we show that DI also suffers from false negatives (FNs) -- an adversary can fool DI by regularising a stolen model's decision boundaries using adversarial training, thereby leading to an FN. To this end, we demonstrate that DI fails to identify a model adversarially trained from a stolen dataset -- the setting where DI is the hardest to evade.

Finally, we discuss the implications of our findings, the viability of fingerprinting-based ownership verification in general, and suggest directions for future work.

****

Client-agnostic Learning and Zero-shot Adaptation for Federated Domain Generalization

Seunghan Yang,Seokeon Choi,Hyunsin Park,Sungha Choi,Simyung Chang,Sungrack Yun

Federated domain generalization (federated DG) aims to learn a client-agnostic global model from various distributed source domains and generalize the model to new clients in completely unseen domains. The main challenges of federated DG are the difficulty of building the global model with local client models from different domains while keeping data private and low generalizability to test clients, where data distribution deviates from those of training clients. To solve these challenges, we present two strategies: (1) client-agnostic learning with mixed instance-global statistics and (2) zero-shot adaptation with estimated statistics. In client-agnostic learning, we first augment local features by using data distribution of other clients via global statistics in the global model's batch normalization layers. This approach allows the generation of diverse domains by mixing local and global feature statistics while keeping data private. Local models then learn client-invariant representations by applying our client-agnostic objectives with the augmented data. Next, we propose a zero-shot adapter to help the learned global model to directly bridge a large domain gap between seen and unseen clients. At inference time, the adapter mixes instance statistics of a test input with global statistics that are vulnerable to distribution shift. With the aid of the adapter, the global model improves generalizability further by reflecting test distribution. We comprehensively evaluate our methods on several benchmarks in federated DG.

****

Towards Robust Model Watermark via Reducing Parametric Vulnerability

Guanhao Gan,Yiming Li,Dongxian Wu,Shu-Tao Xia

Deep neural networks are valuable assets considering their commercial benefits and huge demands for costly annotation and computation resources. To protect the copyright of these deep models, backdoor-based ownership verification becomes popular recently, in which the model owner can watermark the model by embedding a specific behavior before releasing it. The defender (usually the model owner) can identify whether a suspicious third-party model is ``stolen'' from it based on the presence of the behavior. Unfortunately, these watermarks are proven to be vulnerable to removal attacks even like fine-tuning. To further explore this vulnerability, we investigate the parametric space and find there exist many watermark-removed models in the vicinity of the watermarked one, which may be easily used by removal attacks. Inspired by this finding, we propose a minimax formulation to find these watermark-removed models and recover their watermark behavior. Extensive experiments demonstrate that our method improves the robustness of the model watermarking against parametric changes and numerous watermark-removal attacks.

****

This Looks Like It Rather Than That: ProtoKNN For Similarity-Based Classifiers

Yuki Ukai,Tsubasa Hirakawa,Takayoshi Yamashita,Hironobu Fujiyoshi

Among research on the interpretability of deep learning models, the 'this looks like that' framework with ProtoPNet has attracted significant attention. By combining the strong power of deep learning models with the interpretability of case-based inference, ProtoPNet can achieve high accuracy while keeping its reasoning process interpretable. Many methods based on ProtoPNet have emerged to take advantage of this benefit, but despite their practical usefulness, they run into difficulty when utilizing similarity-based classifiers, e.g., in domains where unknown class samples exist. This is because ProtoPNet and its variants adopt the training process specific to linear classifiers, which allows the prototypes to represent useful image features for class recognition. Due to this difficulty, the effectiveness of similarity-based classifiers (e.g., k-nearest neighbor (KNN)) on the 'this looks like that' framework have not been sufficiently examined. To alleviate this problem, we propose ProtoKNN, an extension of ProtoPNet that adopts KNN classifiers. Extensive experiments on multiple open datasets demonstrate that the proposed method can achieve competitive results with a state-of-the-art method.

****

SEQuence-rPPG: A Fast BVP Signal Extraction Method From Frame Sequences

Kegang Wang,Yantao Wei,Mingwen Tong,Jie Gao,ZhongJin Zhao,YuJian Ma,Yi Tian

Non-contact heart rate estimation has essential implications for the development of affective computing and telemedicine. However, existing deep learning-based methods often endeavor to achieve real-time measurements, so a simple, fast, pre-processing-free approach is needed. Our work consists of two main parts. Firstly, we proposed SEQ-rPPG, which first transforms the RGB frame sequence into the original BVP signal sequence by learning-based linear mapping and then outputs the final BVP signal using 1DCNN-based spectral transform, and time-domain filtering. Secondly, to address the shortcomings of the existing dataset in training the model, a new large-scale dataset was collected for training and testing. Our approach achieved competitive results on the collected large dataset(the best) and public dataset UBFC-rPPG(0.81 MAE with 30s time window, test only). It requires no complex pre-processing, has the fastest speed, can run in real-time on mobile ARM CPUs, and can achieve real-time beat-to-beat performance on desktop CPUs. Benefiting from the high-quality training set, other deep learning-based models reduced errors by at least 53$\%$. We compared the methods with and without the spectral transformation, and the results show that the processing in the time domain is effective.

****

Understanding weight-magnitude hyperparameters in training binary networks

Joris Quist,Yunqiang Li,Jan van Gemert

Binary Neural Networks (BNNs) are compact and efficient by using binary weights instead of real-valued weights. Current BNNs use latent real-valued weights during training, where several training hyper-parameters are inherited from real-valued networks. The interpretation of several of these hyperparameters is based on the magnitude of the real-valued weights. For BNNs, however, the magnitude of binary weights is not meaningful, and thus it is unclear what these hyperparameters actually do. One example is weight-decay, which aims to keep the magnitude of real-valued weights small. Other examples are latent weight initialization, the learning rate, and learning rate decay, which influence the magnitude of the real-valued weights. The magnitude is interpretable for real-valued weights, but loses its meaning for binary weights. In this paper we offer a new interpretation of these magnitude-based hyperparameters based on higher-order gradient filtering during network optimization. Our analysis makes it possible to understand how magnitude-based hyperparameters influence the training of binary networks which allows for new optimization filters specifically designed for binary neural networks that are independent of their real-valued interpretation. Moreover, our improved understanding reduces the number of hyperparameters, which in turn eases the hyperparameter tuning effort which may lead to better hyperparameter values for improved accuracy. Code is available at https://github.com/jorisquist/Understanding-WM-HP-in-BNNs

****

Sample-efficient multi-objective molecular optimization with GFlowNets

Yiheng Zhu,Jialu Wu,Chaowen Hu,Jiahuan Yan,Chang-Yu Hsieh,Tingjun Hou,Jian Wu

Many crucial scientific problems involve designing novel molecules with desired properties, which can be formulated as an expensive black-box optimization problem over the discrete chemical space. Computational methods have achieved initial success but still struggle with simultaneously optimizing multiple competing properties in a sample-efficient manner. In this work, we propose a multi-objective Bayesian optimization (MOBO) algorithm leveraging the hypernetwork-based GFlowNets (HN-GFN) as an acquisition function optimizer, with the purpose of sampling a diverse batch of candidate molecular graphs from an approximate Pareto front. Using a single preference-conditioned hypernetwork, HN-GFN learns to explore various trade-offs between objectives. Inspired by reinforcement learning, we further propose a hindsight-like off-policy strategy to share high-performing molecules among different preferences in order to speed up learning for HN-GFN. Through synthetic experiments, we illustrate that HN-GFN has adequate capacity to generalize over preferences. Extensive experiments show that our framework outperforms the best baselines by a large margin in terms of hypervolume in various real-world MOBO settings.

****

Learning Robust Kernel Ensembles with Kernel Average Pooling

Pouya Bashivan,Adam Ibrahim,Amirozhan Dehghani,Yifei Ren

Model ensembles have long been used in machine learning to reduce the variance in individual model predictions, making them more robust to input perturbations. Pseudo-ensemble methods like dropout have also been commonly used in deep learning models to improve generalization. However, the application of these techniques to improve neural networks' robustness against input perturbations remains underexplored. We introduce \emph{Kernel Average Pool (KAP)}, a new neural network building block that applies the mean filter along the kernel dimension of the layer activation tensor. We show that ensembles of kernels with similar functionality naturally emerge in convolutional neural networks equipped with KAP and trained with backpropagation. Moreover, we show that when combined with activation noise, KAP models are remarkably robust against various forms of adversarial attacks. Empirical evaluations on CIFAR10, CIFAR100, TinyImagenet, and Imagenet datasets show substantial improvements in robustness against strong adversarial attacks such as AutoAttack that are on par with adversarially trained networks but are importantly obtained without training on any adversarial examples.

****

Causal Attention to Exploit Transient Emergence of Causal Effect

Xiaolei Ru,Xin-Ya Zhang,Jack Murdoch Moore,Gang Yan

We propose a causal reasoning mechanism called $\textit{causal attention}$ that can improve performance of machine learning models on a class of causal inference tasks by revealing the generation process behind the observed data. We consider the problem of reconstructing causal networks (e.g., biological neural networks) connecting large numbers of variables (e.g., nerve cells), of which evolution is governed by nonlinear dynamics consisting of weak coupling-drive (i.e., causal effect) and strong self-drive (dominants the evolution). The core difficulty is sparseness of causal effect that emerges (the coupling force is significant) only momentarily and otherwise remains dormant in the neural activity sequence. $\textit{Causal attention}$ is designed to guide the model to make inference focusing on the critical regions of time series data where causality may manifest. Specifically, attention coefficients are assigned autonomously by a neural network trained to maximise the Attention-extended Transfer Entropy, which is a novel generalization of the iconic transfer entropy metric. Our results show that, without any prior knowledge of dynamics, $\textit{causal attention}$ explicitly identifies areas where the strength of coupling-drive is distinctly greater than zero. This innovation substantially improves reconstruction performance for both synthetic and real causal networks using data generated by neuronal models widel y used in neuroscience.

****

Imitating Human Behaviour with Diffusion Models

Tim Pearce,Tabish Rashid,Anssi Kanervisto,Dave Bignell,Mingfei Sun,Raluca Georgescu,Sergio Valcarcel Macua,Shan Zheng Tan,Ida Momennejad,Katja Hofmann,Sam Devlin

Diffusion models have emerged as powerful generative models in the text-to-image domain. This paper studies their application as observation-to-action models for imitating human behaviour in sequential environments. Human behaviour is stochastic and multimodal, with structured correlations between action dimensions. Meanwhile, standard modelling choices in behaviour cloning are limited in their expressiveness and may introduce bias into the cloned policy. We begin by pointing out the limitations of these choices. We then propose that diffusion models are an excellent fit for imitating human behaviour, since they learn an expressive distribution over the joint action space. We introduce several innovations to make diffusion models suitable for sequential environments; designing suitable architectures, investigating the role of guidance, and developing reliable sampling strategies. Experimentally, diffusion models closely match human demonstrations in a simulated robotic control task and a modern 3D gaming environment.

****

MetaPhysiCa: Causality-aware Robustness to OOD Initial Conditions in Physics-informed Machine Learning

S Chandra Mouli,Bruno Ribeiro

A fundamental challenge in physics-informed machine learning (PIML) is the design of robust PIML methods for out-of-distribution (OOD) forecasting tasks, where the tasks require learning-to-learn from observations of the same (ODE) dynamical system with different unknown parameters, and demand accurate forecasts even under initial conditions outside the training support. In this work we propose a solution for such tasks, which we define as a meta-learning procedure for causal structural discovery (including invariant risk minimization). Using three different OOD tasks, we empirically observe that the proposed approach significantly outperforms existing state-of-the-art PIML and deep learning methods.

****

Representation Balancing with Decomposed Patterns for Treatment Effect Estimation

Yiyan HUANG,WANG Siyi,Cheuk Hang LEUNG,Qi WU,Dongdong WANG,Zhixiang Huang

Estimating treatment effects from observational data is subject to a problem of covariate shift caused by selection bias. Recent studies have attempted to mitigate this problem by group distance minimization, that is, balancing the distribution of representations between the treated and controlled groups. The rationale behind this is that learning balanced representations while preserving the predictive power of factual outcomes is expected to generalize to counterfactual inference. Inspired by this, we propose a new approach to better capture the patterns that contribute to representation balancing and outcome prediction. Specifically, we derive a theoretical bound that naturally ties the notion of propensity confusion to representation balancing, and further transform the balancing Patterns into Decompositions of Individual propensity confusion and Group distance minimization (PDIG). Moreover, we propose to decompose proxy features into Patterns of Pre-balancing and Balancing Representations (PPBR), as it is insufficient if only balanced representations are considered in outcome prediction. Extensive experiments on simulation and benchmark data confirm not only PDIG leads to mutual reinforcement between individual propensity confusion and group distance minimization, but also PPBR brings improvement to outcome prediction, especially counterfactual inference. We believe these findings are heuristics for further investigation of what affects the generalizability of representation balancing models in counterfactual estimation.

****

Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning

Antonia Creswell,Murray Shanahan,Irina Higgins

Large language models (LLMs) have been shown to be capable of impressive few-shot generalisation to new tasks. However, they still tend to perform poorly on multi-step logical reasoning problems. Here we carry out a comprehensive evaluation of LLMs on 46 tasks that probe different aspects of logical reasoning. We show that language models tend to perform fairly well at single step inference or entailment tasks, but struggle to chain together multiple reasoning steps to solve more complex problems. In light of this, we propose a Selection-Inference (SI) framework that exploits pre-trained LLMs as general processing modules, and alternates between selection and inference to generate a series of interpretable, casual reasoning steps leading to the final answer. We show that a 7B parameter LLM used within the SI framework in a 5-shot generalisation setting, with no fine-tuning, yields a performance improvement of over 100% compared to an equivalent vanilla baseline on a suite of 10 logical reasoning tasks. The same model in the same setting even outperforms a significantly larger 280B parameter baseline on the same suite of tasks. Moreover, answers produced by the SI framework are accompanied by a causal natural-language-based reasoning trace, which has important implications for the safety and trustworthiness of the system.
****

Guided Safe Shooting: model based reinforcement learning with safety constraints
Giuseppe Paolo,Jonas Gonzalez-Billandon,Albert Thomas,Balázs Kégl

In the last decade, reinforcement learning successfully solved complex control tasks and decision-making problems, like the Go board game. Yet, there are few success stories when it comes to deploying those algorithms to real-world scenarios.

One of the reasons is the lack of guarantees when dealing with and avoiding unsafe states, a fundamental requirement in critical control engineering systems. In this paper, we introduce Guided Safe Shooting (GuSS), a model-based RL approach that can learn to control systems with minimal violations of the safety constraints. The model is learned on the data collected during the operation of the system in an iterated batch fashion, and is then used to plan for the best action to perform at each time step. We propose three different safe planners, one based on a simple random shooting strategy and two based on MAP-Elites, a more advanced divergent-search algorithm. Experiments show that these planners help the learning agent avoid unsafe situations while maximally exploring the state space, a necessary aspect when learning an accurate model of the system. Furthermore, compared to model-free approaches, learning a model allows GuSS reducing the number of interactions with the real-system while still reaching high rewards, a fundamental requirement when handling engineering systems.
****

Contrastive Meta-Learning for Partially Observable Few-Shot Learning
Adam Jelley,Amos Storkey,Antreas Antoniou,Sam Devlin

Many contrastive and meta-learning approaches learn representations by identifying common features in multiple views. However, the formalism for these approaches generally assumes features to be shared across views to be captured coherently. We consider the problem of learning a unified representation from partial observations, where useful features may be present in only some of the views. We approach this through a probabilistic formalism enabling views to map to representations with different levels of uncertainty in different components; these views can then be integrated with one another through marginalisation over that uncertainty. Our approach, Partial Observation Experts Modelling (POEM), then enables us to meta-learn consistent representations from partial observations. We evaluate our approach on an adaptation of a comprehensive few-shot learning benchmark, Meta-Dataset, and demonstrate the benefits of POEM over other meta-learning methods at representation learning from partial observations. We further demonstrate the utility of POEM by meta-learning to represent an environment from partial views observed by an agent exploring the environment.

****

Enhancing the Inductive Biases of Graph Neural ODE for Modeling Physical Systems
Suresh Bishnoi,Ravinder Bhattoo,Jayadeva Jayadeva,Sayan Ranu,N M Anoop Krishnan

Neural networks with physics-based inductive biases such as Lagrangian neural networks (LNNs), and Hamiltonian neural networks (HNNs) learn the dynamics of physical systems by encoding strong inductive biases. Alternatively, Neural ODEs with appropriate inductive biases have also been shown to give similar performances. However, these models, when applied to particle-based systems, are transductive in nature and hence, do not generalize to large system sizes. In this paper, we present a graph-based neural ODE, GNODE, to learn the time evolution of dynamical systems. Further, we carefully analyze the role of different inductive biases on the performance of GNODE. We show that similar to LNN and HNN, encoding the constraints explicitly can significantly improve the training efficiency and performance of GNODE significantly. Our experiments also assess the value of additional inductive biases, such as Newton’s third law, on the final performance of the model. We demonstrate that inducing these biases can enhance the performance of the model by orders of magnitude in terms of both energy violation and rollout error. Interestingly, we observe that the GNODE trained with the most effective inductive biases, namely MCGNODE, outperforms the graph versions of LNN and HNN, namely, Lagrangian graph networks (LGN) and Hamiltonian graph networks (HGN) in terms of energy violation error by ~4 orders of magnitude for a pendulum system, and ~2 orders of magnitude for spring systems. These results suggest that NODE-based systems can give competitive performances with energy-conserving neural networks by employing appropriate inductive biases.

****

Efficient Planning in a Compact Latent Action Space
zhengyao jiang,Tianjun Zhang,Michael Janner,Yueying Li,Tim Rocktäschel,Edward Grefenstette,Yuandong Tian

Planning-based reinforcement learning has shown strong performance in tasks in discrete and low-dimensional continuous action spaces. However, planning usually brings significant computational overhead for decision making, so scaling such methods to high-dimensional action spaces remains challenging. To advance efficient planning for high-dimensional continuous control, we propose Trajectory Autoencoding Planner (TAP), which learns low-dimensional latent action codes with a state-conditional VQ-VAE. The decoder of the VQ-VAE thus serves as a novel dynamics model that takes latent actions and current state as input and reconstructs long-horizon trajectories. During inference time, given a starting state, TAP searches over discrete latent actions to find trajectories that have both high probability under the training distribution and high predicted cumulative reward. Empirical evaluation in the offline RL setting demonstrates low decision latency which is indifferent to the growing raw action dimensionality. For Adroit robotic hand manipulation tasks with high-dimensional continuous action space, TAP surpasses existing model-based methods by a large margin and also beats strong model-free actor-critic baselines.

****

Improved Stein Variational Gradient Descent with Importance Weights
Lukang Sun,Peter Richtárik

Stein Variational Gradient Descent~(\algname{SVGD}) is a popular sampling algorithm used in various machine learning tasks. It is well known that \algname{SVGD} arises from a discretization of the kernelized gradient flow of the Kullback-Leibler divergence $\KL\left(\cdot\mid\pi\right)$, where π is the target distribution. In this work, we propose to enhance \algname{SVGD} via the introduction of {\em importance weights}, which leads to a new method for which we coin the name \algname{β-SVGD}. In the continuous time and infinite particles regime, the time for this flow to converge to the equilibrium distribution π, quantified by the Stein Fisher information, depends on ρ_0 and π very we akly. This is very different from the kernelized gradient flow of Kullback-Leibler divergence, whose time complexity depends on $\KL\left(\rho_0\mid\pi\right)$. Under certain assumptions, we provide a descent lemma for the population limit \algname{β-SVGD}, which covers the descent lemma for the population limit \algname{SVGD} when $\beta\to 0$. We also illustrate the advantages of \algname{β-SVGD} over \algname{SVGD} by simple experiments.

****

Correlative Information Maximization Based Biologically Plausible Neural Networks for Correlated Source Separation

Bariscan Bozkurt,Ate■ ■sfendiyaro■lu,Cengiz Pehlevan,Alper Tunga Erdogan

The brain effortlessly extracts latent causes of stimuli, but how it does this at the network level remains unknown. Most prior attempts at this problem proposed neural networks that implement independent component analysis, which works under the limitation that latent elements are mutually independent. Here, we relax this limitation and propose a biologically plausible neural network that extracts correlated latent sources by exploiting information about their domains. To derive this network, we choose maximum correlative information transfer from inputs to outputs as the separation objective under the constraint that the outputs are restricted to their presumed sets. The online formulation of this optimization problem naturally leads to neural networks with local learning rules. Our framework incorporates infinitely many source domain choices and flexibly models complex latent structures. Choices of simplex or polytopic source domains result in networks with piecewise-linear activation functions. We provide numerical examples to demonstrate the superior correlated source separation capability for both synthetic and natural sources.

****

Simplicity bias leads to amplified performance disparities

Samuel Bell,Levent Sagun

The simple idea that not all things are equally difficult has surprising implications when applied in a fairness context.

In this work we explore how "difficulty" is model-specific, such that different models find different parts of a dataset challenging.

When difficulty correlates with group information, we term this difficulty disparity.

Drawing a connection with recent work exploring the inductive bias towards simplicity of SGD-trained models, we show that when such a disparity exists, it is further amplified by commonly-used models.

We quantify this amplification factor across a range of settings aiming towards a fuller understanding of the role of model bias. We also present a challenge to the simplifying assumption that ``fixing'' a dataset is sufficient to ensure unbiased performance.

****

Annealed Fisher Implicit Sampler

Weijian Luo,Boya Zhang,Zhihua Zhang

Sampling from an un-normalized target distribution is an important problem in many scientific fields. An implicit sampler uses a parametric transform $x=G_\theta(z)$ to push forward an easy-to-sample latent code z to obtain a sample x. Such samplers are favored for fast inference speed and flexible architecture. Thus it is appealing to train an implicit sampler for sampling from the un-normalized target. In this paper, we propose a novel approach to training an implicit sampler by minimizing the Fisher Divergence between sampler and target distribution. We find that the trained sampler works well for relatively simple targets but may fail for more complicated multi-modal targets. To improve the training for multi-modal targets, we propose another adaptive training approach that trains the sampler to gradually learn a sequence of annealed distributions. We construct the annealed distribution path to bridge a simple distribution and the complicated target. With the annealed approach, the sampler is capable of handling challenging multi-modal targets. In addition, we also introduce a few MCMC correction steps after the sampler to better spread the samples. We call our proposed sampler \emph{the Annealed Fisher Implicit Sampler} (AFIS). We test AFIS on several sampling benchmarks. The experiments show that our AFIS outperforms baseline methods in many aspects. We also show in theory that the added MC correction steps get faster mixing by using the learned sampler as MCMC's initialization.

****

Do You Remember? Overcoming Catastrophic Forgetting for Fake Audio Detection

XiaoHui Zhang,Jiangyan Yi,Chenglong Wang,Chu Yuan Zhang,Jianhua Tao

Current fake audio detection algorithms achieve promising performances on most datasets. However, their performance may be significantly degraded when dealing with audio of a different dataset. The orthogonal weight modification to overcome catastrophic forgetting does not consider the similarity of some audio, including fake audio obtained by the same algorithm and genuine audio, on different datasets. To overcome this limitation, we propose a continual learning algorithm for fake audio detection to overcome catastrophic forgetting, called Regularized Adaptive Weight Modification (RAWM). Specifically, when fine-tuning a detection network, our approach adaptively computes the direction of weight modification according to the ratio of genuine utterances and fake utterances. The adaptive modification direction ensures the network can detect fake audio on the new dataset while preserving its knowledge of previous model, thus mitigating catastrophic forgetting. In addition, orthogonal weight modification of fake audios in the new dataset will skew the distribution of inferences on audio in the previous dataset with similar acoustic characteristics, so we introduce a regularization constraint to force the network to remember this distribution. We evaluate our approach across multiple datasets and obtain a significant performance improvement on cross-dataset experiments.

****

Towards Conditionally Dependent Masked Language Models

Lucas Torroba Hennigen,Yoon Kim

Masked language modeling has proven to be an effective paradigm for learning representations of language. However, when multiple tokens are masked out, the masked language model's (MLM) distribution over the masked positions assumes that the masked tokens are conditionally independent given the unmasked tokens---an assumption that does not hold in practice. Existing work addresses this limitation by interpreting the sum of unary scores (i.e., the logits or the log probabilities of single tokens when conditioned on all others) as the log potential a Markov random field (MRF). While this new model no longer makes any independence assumptions, it remains unclear whether this approach (i) results in a good probabilistic model of language and further (ii) derives a model that is faithful (i.e., has matching unary distributions) to the original model. This paper studies MRFs derived this way in a controlled setting where only two tokens are masked out at a time, which makes it possible to compute exact distributional properties. We find that such pairwise MRFs are often worse probabilistic models of language from a perplexity standpoint, and moreover have unary distributions that do not match the unary distributions of the original MLM. We then study a statistically-motivated iterative optimization algorithm for deriving joint pairwise distributions that are more compatible with the original unary distributions. While this iterative approach outperforms the MRF approach, the algorithm itself is too expensive to be practical. We thus amortize this optimization process through a parameterized feed-forward layer that learns to modify the original MLM's pairwise distributions to be both non-independent and faithful, and find that this approach outperforms the MLM for scoring pairwise tokens.

****

Leveraging Importance Weights in Subset Selection

Gui Citovsky,Giulia DeSalvo,Sanjiv Kumar,Srikumar Ramalingam,Afshin Rostamizadeh,Yunjuan Wang

We present a subset selection algorithm designed to work with arbitrary model families in a practical batch setting. In such a setting, an algorithm can sample examples one at a time but, in order to limit overhead costs, is only able to update its state (i.e. further train model weights) once a large enough batch of examples is selected. Our algorithm, IWeS, selects examples by importance sampli ng where the sampling probability assigned to each example is based on the entropy of models trained on previously selected batches. IWeS admits significant performance improvement compared to other subset selection algorithms for seven publicly available datasets. Additionally, it is competitive in an active learning setting, where the label information is not available at selection time. We also provide an initial theoretical analysis to support our importance weighting approach, proving generalization and sampling rate bounds.

****

Interactive Sequential Generative Models

Dennis Fassmeyer,Pascal Fassmeyer,Ulf Brefeld

Understanding spatiotemporal relationships among several agents is of considerable relevance for many domains. Team sports represent a particularly interesting real-world proving ground since modeling interacting athletes requires capturing highly dynamic and complex agent-agent dependencies in addition to temporal components. However, existing generative methods in this field either entangle all latent factors into a single variable and are thus constrained in practical applicability, or they focus on uncovering interaction structures, which restricts their generative ability. To address this gap, we propose a framework for multiagent trajectories that augments sequential generative models with latent social structures. First, we derive a novel objective via approximate inference using a disentangled latent space that accurately describes the data generating process in such systems. Based on the proposed training criterion, we then present a model architecture that unifies insights from neural interaction inference and graph-structured variational recurrent neural networks for generating collective movements while allocating latent information. We validate our model on data from professional soccer and basketball. Our framework not only improves upon existing state-of-the-art approaches in forecasting trajectories, but also infers semantically meaningful representations that can be used in downstream tasks.

****

Gradient flow in the gaussian covariate model: exact solution of learning curves and multiple descent structures

Antoine Bodin,Nicolas Macris

A recent line of work has shown remarkable behaviors of the generalization error curves in simple learning models. Even the least-squares regression has shown atypical features such as the model-wise double descent, and further works have observed triple or multiple descents. Another important characteristic are the epoch-wise descent structures which emerge during training. The observations of model-wise and epoch-wise descents have been analytically derived in limited theoretical settings (such as the random feature model) and are otherwise experimental. In this work, we provide a full and unified analysis of the whole time-evolution of the generalization curve, in the asymptotic large-dimensional regime and under gradient-flow, within a wider theoretical setting stemming from a gaussian covariate model. In particular, we cover most cases already disparately observed in the literature, and also provide examples of the existence of multiple descent structures as a function of a model parameter or time. Furthermore, we show that our theoretical predictions adequately match the learning curves obtained by gradient descent over realistic datasets.

Technically we compute averages of rational expressions involving random matrices using recent developments in random matrix theory based on "linear pencils". Another contribution, which is also of independent interest in random matrix theory, is a new derivation of related fixed point equations (and an extension there-off) using Dyson brownian motions.

****

Copy is All You Need

Tian Lan,Deng Cai,Yan Wang,Heyan Huang,Xian-Ling Mao

The dominant text generation models compose the output by sequentially selecting words from a fixed vocabulary. In this paper, we formulate text generation as progressively copying text segments (e.g., words or phrases) from an existing text collection. We compute the contextualized representations of meaningful text segments and index them using efficient vector search toolkits. The task of text generation is then decomposed into a series of copy-and-paste operations: at each time step, we seek suitable text spans from the text collection rather than selecting from a standalone vocabulary. Experiments on the standard language modeling benchmark (WikiText-103) show that our approach achieves better generation quality according to both automatic and human evaluations. Besides, its inference efficiency is comparable to token-level autoregressive models thanks to the reduction of decoding steps. We also show that our approach allows for effective domain adaptation by simply switching to domain-specific text collection without extra training. Finally, we observe that our approach attains additional performance gains by simply scaling up to larger text collections, again without further training.\footnote{Our source codes are publicly available at \url{https://github.com/gmftbyGMFTBY/Copyisallyouneed}.}

****

Graph Backup: Data Efficient Backup Exploiting Markovian Transitions

zhengyao jiang,Tianjun Zhang,Robert Kirk,Tim Rocktäschel,Edward Grefenstette

The successes of deep Reinforcement Learning (RL) are limited to settings where we have a large stream of online experiences, but applying RL in the data-efficient setting with limited access to online interactions is still challenging. A key to data-efficient RL is good value estimation, but current methods in this space fail to fully utilise the structure of the trajectory data gathered from the environment. In this paper, we treat the transition data of the MDP as a graph, and define a novel backup operator, Graph Backup, which exploits this graph structure for better value estimation. Compared to multi-step backup methods such as n-step Q-Learning and TD(λ), Graph Backup can perform counterfactual credit assignment and gives stable value estimates for a state regardless of which trajectory the state is sampled from. Our method, when combined with popular off-policy value-based methods, provides improved performance over one-step and multi-step methods on a suite of data-efficient RL benchmarks including MiniGrid, Minatar and Atari100K. We further analyse the reasons for this performance boost through a novel visualisation of the transition graphs of Atari games.

****

Linearised Implicit Variational Inference

Anshuk Uppal,Wouter Boomsma,Jes Frellsen

Bayesian neural networks (BNNs) are touted for robustness under data drift, resilience to overfitting and catastrophic forgetting whilst also producing actionable uncertainty estimates. In variational inference, these elegant properties are contingent on the expressivity of the variational approximation. Posteriors over parameters of large models are usually multimodal and highly correlated and hence cannot be well-approximated by simple, prescribed densities. We posit implicit variational distributions specified using differentiable generators are more flexible and propose a novel bound for training BNNs using such approximations (amortized neural samplers). The proposed bound uses an approximation of the variational distribution's entropy by locally linearising the generator. Unlike existing works, our method does not require a discriminator network and moves away from an unfavourable adversarial objective. Our formulation resembles normalizing flows but does not necessitate invertibility of the generator. Moreover, we use a differentiable numerical lower bound on the Jacobians of the generator, mitigating computational concerns. We report log-likelihoods on UCI datasets competitive with deep ensembles and test our method on out-of-distribution benchmarks.

****

Adversarial Driving Policy Learning by Misunderstanding the Traffic Flow

Dongkun Zhang,Jintao Xue,Jingke Wang,Yunkai Wang,Yuxiang Cui,Eryun Liu,Wei Jing,Rong Xiong,Junbo Chen,Yue Wang

Acquiring driving policies that can transfer to unseen environments is essential for driving in dense traffic flows. Adversarial training is a promising path to improve robustness under disturbances. Most prior works leverage few agents to induce driving policy's failures. However, we argue that directly implementing this training framework into dense traffic flow degrades transferability in unseen environments. In this paper, we propose a novel robust policy training framework that is capable of applying adversarial training based on a coordinated traff ic flow. We start by building up a coordinated traffic flow where agents are allowed to communicate Social Value Orientations (SVOs). Adversary emerges when the traffic flow misunderstands the SVO of driving agent. We utilize this property to formulate a minimax optimization problem where the driving policy maximizes its own reward and a spurious adversarial policy minimizes it. Experiments demonstrate that our adversarial training framework significantly improves zero-shot transfer performance of the driving policy in dense traffic flows compared to existing algorithms.

****

Differentiable and transportable structure learning

Jeroen Berrevoets,Nabeel Seedat,Fergus Imrie,Mihaela van der Schaar

Directed acyclic graphs (DAGs) encode a lot of information about a particular distribution in its structure. However, compute required to infer these structures is typically super-exponential in the number of variables, as inference requires a sweep of a combinatorially large space of potential structures. That is, until recent advances made it possible to search this space using a differentiable metric, drastically reducing search time. While this technique– named NOTEARS –is widely considered a seminal work in DAG-discovery, it concedes an important property in favour of differentiability: transportability. To be transportable, the structures discovered on one dataset must apply to another dataset from the same domain. In our paper, we introduce D-Struct which recovers transportability in the discovered structures through a novel architecture and loss function, while remaining completely differentiable. Because D-Struct remains differentiable, our method can be easily adopted in existing differentiable architectures, as was previously done with NOTEARS. In our experiments, we empirically validate D-Struct with respect to edge accuracy and structural Hamming distance in a variety of settings.

****

Distributed Inference and Fine-tuning of Large Language Models Over The Internet

Alexander Borzunov,Dmitry Baranchuk,Tim Dettmers,Max Ryabinin,Younes Belkada,Artem Chumachenko,Pavel Samygin,Colin Raffel

Large language models (LLMs) are useful in many NLP tasks and become more capable with size, scaling to over 100 billion parameters. With the release of BLOOM-176B and OPT-175B, everyone can download pretrained models of this scale. Still, using a pre-trained 100B+ model requires high-end hardware, making it inaccessible to most researchers. Recent studies in memory-efficient training (e.g. offloading) could alleviate these costs, but they do not cover important use cases of LLMs, such as autoregressive inference. In this work, we investigate methods for cost-efficient inference of large language models, comparing local and distributed strategies. We observe that a large enough model (100B+) could run efficiently on geodistributed devices in a consumer-grade network, for example by connecting existing compute resources of multiple research groups or pooling under-utilized compute from multiple cloud regions. To run LLMs in this unconventional setting, we develop a fault-tolerant algorithm for inferencing language models. We propose Petals - a decentralized system for running LLMs - and show that it can run BLOOM-176B over the Internet over $10\times$ faster than offloading for sequential generation. We evaluate the performance of our system in both simulated conditions and an actual distributed system spanning two continents. The design of Petals allows participants to inference, and fine-tune, or inference fine-tuned models simultaneously without affecting each other's results.

****

Association Rules in QUBO Samples and Where to Find Them

Tian Huang,Siong Thye Goh,Rick Siow Mong Goh,Tao Luo

There are sometimes strong associations between variables in the samples to a Quadratic Unconstrained Binary Optimization (QUBO) problem. A natural question arises to us: Are there any value in these association? We study max-cut problem and observe that association can be represented as rules to simplify QUBO problem. Classical and quantum annealers work better when the problem size is smaller. To effectively and efficiently find associations between variables, we adapt traditional association rule mining in the case of QUBO samples and propose a Fast A ssociation Rule Mining algorithm (FARM) specifically for mining QUBO samples. We also propose strategies and a workflow to select and apply promising rules and simplify QUBO problems. We evaluate our method on D-Wave Quantum Annealer as well as Fujitsu Digital Annealer. The experiments demonstrate the utility of FARM as a visualisation tool for understanding associations in QUBO samples. The results also demonstrate the potential of our method in closing the gap between samples and ground truth. The source code will be disclosed to the public if the manuscript is accepted.

****

Why adversarial training can hurt robust accuracy

Jacob Clarysse,Julia Hörrmann,Fanny Yang

Machine learning classifiers with high test accuracy often perform poorly under adversarial attacks. It is commonly believed that adversarial training alleviates this issue. In this paper, we demonstrate that, surprisingly, the opposite can be true for a natural class of perceptible perturbations --- even though adversarial training helps when enough data is available, it may in fact hurt robust generalization in the small sample size regime. We first prove this phenomenon for a high-dimensional linear classification setting with noiseless observations. Using intuitive insights from the proof, we could surprisingly find perturbations on standard image datasets for which this behavior persists. Specifically, it occurs for perceptible attacks that effectively reduce class information such as object occlusions or corruptions.

****

ExpressivE: A Spatio-Functional Embedding For Knowledge Graph Completion

Aleksandar Pavlovic,Emanuel Sallinger

Knowledge graphs are inherently incomplete. Therefore substantial research has been directed toward knowledge graph completion (KGC), i.e., predicting missing triples from the information represented in the knowledge graph (KG). KG embedding models (KGEs) have yielded promising results for KGC, yet any current KGE is incapable of: (1) fully capturing vital inference patterns (e.g., composition), (2) capturing prominent patterns jointly (e.g., hierarchy and composition), and (3) providing an intuitive interpretation of captured patterns. In this work, we propose ExpressivE, a fully expressive spatio-functional KGE that solves all these challenges simultaneously. ExpressivE embeds pairs of entities as points and relations as hyper-parallelograms in the virtual triple space $\mathbb{R}^{2d}$. This model design allows ExpressivE not only to capture a rich set of inference patterns jointly but additionally to display any supported inference pattern through the spatial relation of hyper-parallelograms, offering an intuitive and consistent geometric interpretation of ExpressivE embeddings and their captured patterns. Experimental results on standard KGC benchmarks reveal that ExpressivE is competitive with state-of-the-art KGEs and even significantly outperforms them on WN18RR.

****

FedPD: Defying data heterogeneity through privacy distillation

Zhiqin Brian Yang,Yonggang Zhang,Yu Zheng,Zhenheng TANG,Xiaowen Chu,Hao Peng,Bo Han

Model performance of federated learning (FL) typically suffers from data heterogeneity, i.e., data distribution varies with clients. Advanced works have already shown great potential for sharing client information to mitigate data heterogeneity. Yet, some literature shows a dilemma in preserving strong privacy and promoting model performance simultaneously. Revisiting the purpose of sharing information motivates us to raise the fundamental questions: Which part of the data is more critical for model generalization? Which part of the data is more privacy-sensitive? Can we solve this dilemma by sharing useful (for generalization) features and maintaining more sensitive data locally? Our work sheds light on data-dominated sharing and training, in a way that we decouple original training data into sensitive features and generalizable features. To be specific, we propose a \textbf{Fed}erated \textbf{P}rivacy \textbf{D}istillation framework named FedPD to alleviate the privacy-performance dilemma. Namely, FedPD keeps the distilled sensitive features locally and constructs a global dataset using shared general izable features in a differentially private manner. Accordingly, clients can perform local training on both the local and securely shared data for acquiring high model performance and avoiding the leakage of not distilled privacy. Theoretically, we demonstrate the superiority of the sharing-only useful feature strategy over sharing raw data. Empirically, we show the efficacy of FedPD in promoting performance with comprehensive experiments.

****

Harnessing Client Drift with Decoupled Gradient Dissimilarity

Zhenheng TANG,Yonggang Zhang,Shaohuai Shi,Xinmei Tian,Tongliang Liu,Bo Han,Xiaowen Chu

The performance of Federated learning (FL) typically suffers from client drift caused by heterogeneous data, where data distributions vary with clients. Recent studies show that the gradient dissimilarity between clients induced by the data distribution discrepancy causes the client drift. Thus, existing methods mainly focus on correcting the gradients. However, it is challenging to identify which client should (or not) be corrected. This challenge raises a series of questions: will the local training, without gradient correction, contribute to the server model's generalization of other clients' distributions? when the generalization contribution holds? how to address the challenge when it fails? To answer these questions, we analyze the generalization contribution of local training and conclude that the generalization contribution of local training is bounded by the conditional Wasserstein distance between clients' distributions. Thus, the key to promote generalization contribution is to leverage similar conditional distributions for local training. As collecting data distribution can cause privacy leakage, we propose decoupling the deep models, i.e., splitting into high-level models and low-level models, for harnessing client drift. Namely, high-level models are trained on shared feature distributions, causing promoted generalization contribution and alleviated gradient dissimilarity. Experimental results demonstrate that FL with decoupled gradient dissimilarity is robust to data heterogeneity.

****

SeKron: A Decomposition Method Supporting Many Factorization Structures

Marawan Gamal,Ali Mosleh,Marzieh S. Tahaei,Vahid Partovi Nia

While convolutional neural networks (CNNs) have become the de facto standard for most image processing and computer vision applications, their deployment on edge devices remains challenging. Tensor decomposition methods provide a means of compressing CNNs to meet the wide range of device constraints by imposing certain factorization structures on their convolution tensors. However, being limited to the small set of factorization structures presented by state-of-the-art decomposition approaches can lead to sub-optimal performance. We propose SeKron, a novel tensor decomposition method that offers a wide variety of factorization structures, using sequences of Kronecker products. By recursively finding approximating Kronecker factors, we arrive at optimal decompositions for each of the factorization structures. We show that SeKron is a flexible decomposition that generalizes widely used methods, such as Tensor-Train (TT), Tensor-Ring (TR), Canonical Polyadic (CP) and Tucker decompositions. Crucially, we derive an efficient convolution projection algorithm shared by all SeKron structures, leading to seamless compression of CNN models. We validate SeKron for model compression on both high-level and low-level computer vision tasks and find that it outperforms state-of-the-art decomposition methods.

****

Localized Randomized Smoothing for Collective Robustness Certification

Jan Schuchardt,Tom Wollschläger,Aleksandar Bojchevski,Stephan Günnemann

Models for image segmentation, node classification and many other tasks map a single input to multiple labels. By perturbing this single shared input (e.g. the image) an adversary can manipulate several predictions (e.g. misclassify several pixels). Collective robustness certification is the task of provably bounding the number of robust predictions under this threat model. The only dedicated method that goes beyond certifying each output independently is limited to strictly local models, where each prediction is associated with a small receptive field.

We propose a more general collective robustness certificate for all types of models. We further show that this approach is beneficial for the larger class of softly local models, where each output is dependent on the entire input but assigns different levels of importance to different input regions (e.g. based on their proximity in the image). The certificate is based on our novel localized randomized smoothing approach, where the random perturbation strength for different input regions is proportional to their importance for the outputs. Localized smoothing Pareto-dominates existing certificates on both image segmentation and node classification tasks, simultaneously offering higher accuracy and stronger certificates.

****

Learning Dictionaries over Datasets through Wasserstein Barycenters

Eduardo Fernandes Montesuma,Fred Maurice Ngole Mboula,Antoine Souloumiac

Dictionary learning consists of trying to represent objects in terms of basic elements (atoms) weighted by an importance factor (representation). Non-linear dictionary learning using optimal transport as a metric has been previously studied for normalized non-negative data on a fixed grid. We propose a new framework by using Wasserstein Dictionary Learning on datasets understood as empirical distributions. We leverage Wasserstein barycenters for learning a dictionary of virtual datasets and embeddings in a simplex. We apply our method for unsupervised domain adaptation, improving the state-of-the-art over 1.96% and 2.70%, respectively, and manifold learning of Gaussian distributions and color histograms.

****

Learning Interpretable Neural Discrete Representation for Time Series Classification

Etienne Le Naour,Ghislain Agoua,Nicolas Baskiotis,Vincent Guigue

Time series classification is a challenging research field with many real-life applications. Recent advances in deep learning have significantly improved the state of the art: recurrent or convolutional architectures allow automatic extraction of complex discriminating patterns that improve performance. Those approaches suffer from a lack of interpretability: the patterns are mapped into a high dimensional latent vector space, they are not representable in the time domain, and are often even not localizable. In this paper, we present a novel neural convolutional architecture that aims to provide a trade-off between interpretability and effectiveness based on the learning of a dictionary of discrete representations. The proposed model guarantees (1) that a small number of patterns are learned, and they are visualizable and interpretable (2) a shift equivariance property of the model associated with a time-consistency of the representation (3) a linear classifier over a limited number of patterns leading to an explainable decision. To ensure the robustness of the discrete representation, they are learned in an unsupervised process independently of the classification task. This allows further great performances in transfer learning. We present extensive experiments on the UCR benchmark wrt usual baselines. The interpretability of the model is illustrated empirically. The chosen trade-off results obviously in a decrease in performance compared to the state of the art. The performance drop is however limited and very dependent on the application domain. The experiments highlight the efficiency of the model for the transfer learning task, showing the robustness of the representations.

****

Representational Dissimilarity Metric Spaces for Stochastic Neural Networks

Lyndon Duong,Jingyang Zhou,Josue Nassar,Jules Berman,Jeroen Olieslagers,Alex H Williams

Quantifying similarity between neural representations---e.g. hidden layer activation vectors---is a perennial problem in deep learning and neuroscience research. Existing methods compare deterministic responses (e.g. artificial networks that lack stochastic layers) or averaged responses (e.g., trial-averaged firing rates in biological data). However, these measures of _deterministic_ representational similarity ignore the scale and geometric structure of noise, both of which play important roles in neural computation. To rectify this, we generalize previously proposed shape metrics (Williams et al. 2021) to quantify differences in _

stochastic_ representations. These new distances satisfy the triangle inequality, and thus can be used as a rigorous basis for many supervised and unsupervised analyses. Leveraging this novel framework, we find that the stochastic geometries of neurobiological representations of oriented visual gratings and naturalistic scenes respectively resemble untrained and trained deep network representations. Further, we are able to more accurately predict certain network attributes (e.g. training hyperparameters) from its position in stochastic (versus deterministic) shape space.
****
Hierarchical Prototypes for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning
Jiaxian Guo,Mingming Gong,Yali Du,Zhen Wang,Dacheng Tao
By incorporating the environment-specific factor into the dynamics prediction, model-based reinforcement learning (MBRL) is able to generalise to environments with diverse dynamics.In the majority of real-world scenarios, the environment-specific factor is not observable, so existing methods attempt to estimate it from historical transition segments. Nevertheless,earlier research was unable to identify distinct clusters for environment-specific factors learned from different environments, resulting in poor performance.
To address this issue,
We introduce a set of environmental prototypes to represent the environmental-specified representation for each environment. By encouraging learned environment-specific factors to resemble their assigned environmental prototypes more closely, the discrimination between factors estimated from distinct environments will be enhanced. To learn such prototypes, we first construct prototypes for each sampled trajectory and then hierarchically combine trajectory prototypes with similar semantics into one environmental prototype. Experiments demonstrate that environment-specific factors estimated by our method have superior clustering performance and can consistently improve MBRL's generalisation performance in six environments consistently.
****
Irregularity Reflection Neural Network for Time Series Forecasting
YoungJun Choo,Adrian Matias Chung Baek,Namhun Kim,Han-Gyun Woo
Time series forecasting is a long-standing challenge in a variety of industries, and deep learning stands as the mainstream paradigm for handling this forecasting problem. With recent success, representations of time series components (e.g., trend and seasonality) are also considered in the learning process of the models. However, the residual remains under explored due to difficulty in formulating its inherent complexity. In this study, we propose a novel Irregularity Reflection Neural Network (IRN) that reflect the residual for the time series forecasting. First, we redefine the residual as the irregularity and express it as a sum of individual, short regular waves considering the Fourier series in a micro perspective. Second, we design a module, based on the convolutional architectures to mimic the variables of the derived irregularity representation, named Irregularity Representation Block (IRB). IRN comprises IRB on top of a forecasting model to learn the irregularity representation of time series. Extensive experiments on multiple real-world datasets demonstrate that IRN outperforms the state-of-the-art benchmarks in time series forecasting tasks.
****
Sequential Learning of Neural Networks for Prequential MDL
Jorg Bornschein,Yazhe Li,Marcus Hutter
Minimum Description Length (MDL) provides a framework and an objective for principled model evaluation. It formalizes Occam's Razor and can be applied to data from non-stationary sources. In the prequential formulation of MDL, the objective is to minimize the cumulative next-step log-loss when sequentially going through the data and using previous observations for parameter estimation. It thus closely resembles a continual- or online-learning problem. In this study, we evaluate approaches for computing prequential description lengths for image classification datasets with neural networks. Considering the computational cost, we find that online-learning with rehearsal has favorable performance compared to the pr

eviously widely used block-wise estimation. We propose forward-calibration to better align the models predictions with the empirical observations and introduce replay-streams, a minibatch incremental training technique to efficiently implement approximate random replay while avoiding large in-memory replay buffers. As a result, we present description lengths for a suite of image classification datasets that improve upon previously reported results by large margins.

****

Relaxed Attention for Transformer Models

Timo Lohrenz,Björn Möller,Zhengyang Li,Tim Fingscheidt

The powerful modeling capabilities of all-attention-based transformer architectures often cause overfitting and - for natural language processing tasks - lead to an implicitly learned internal language model in the autoregressive transformer decoder complicating the integration of external language models. In this paper, we explore relaxed attention, a simple and easy-to-implement smoothing of the attention weights, yielding a two-fold improvement to the general transformer architecture: First, relaxed attention provides regularization when applied to the self-attention layers in the encoder. Second, we show that it naturally supports the integration of an external language model as it suppresses the implicitly learned internal language model by relaxing the cross attention in the decoder. We demonstrate the benefit of relaxed attention across several tasks with clear improvement in combination with recent benchmark approaches. Specifically, we exceed the former state-of-the-art performance of 26.90% word error rate on the largest public lip-reading LRS3 benchmark with a word error rate of 26.31%, as well as we achieve a top-performing BLEU score of 37.67 on the IWSLT14 (DE$\rightarrow$EN) machine translation task without external language models and virtually no additional model parameters. Code and models will be made publicly available.

****

SynBench: Task-Agnostic Benchmarking of Pretrained Representations using Synthetic Data

Ching-Yun Ko,Pin-Yu Chen,Jeet Mohapatra,Payel Das,Luca Daniel

Recent success in fine-tuning large models, that are pretrained on broad data at scale, on downstream tasks has led to a significant paradigm shift in deep learning, from task-centric model design to task-agnostic representation learning and task-specific fine-tuning. As the representations of pretrained models are used as a foundation for different downstream tasks, this paper proposes a new task-agnostic framework, \textit{SynBench}, to measure the quality of pretrained representations using synthetic data. We set up a reference by a theoretically-derived robustness-accuracy tradeoff of the class conditional Gaussian mixture. Given a pretrained model, the representations of data synthesized from the Gaussian mixture are used to compare with our reference to infer the quality. By comparing the ratio of area-under-curve between the raw data and their representations, SynBench offers a quantifiable score for robustness-accuracy performance benchmarking. Our framework applies to a wide range of pretrained models taking continuous data inputs and is independent of the downstream tasks and datasets. Evaluated with several pretrained vision transformer models, the experimental results show that our SynBench score well matches the actual linear probing performance of the pre-trained model when fine-tuned on downstream tasks. Moreover, our framework can be used to inform the design of robust linear probing on pretrained representations to mitigate the robustness-accuracy tradeoff in downstream tasks.

****

Learning topology-preserving data representations

Ilya Trofimov,Daniil Cherniavskii,Eduard Tulchinskii,Nikita Balabin,Evgeny Burnaev,Serguei Barannikov

We propose a method for learning topology-preserving data representations (dimensionality reduction). The method aims to provide topological similarity between the data manifold and its latent representation via enforcing the similarity in topological features (clusters, loops, 2D voids, etc.) and their localization. The core of the method is the minimization of the Representation Topology Divergence (RTD) between original high-dimensional data and low-dimensional representa

tion in latent space. RTD minimization provides closeness in topological features with strong theoretical guarantees.
We develop a scheme for RTD differentiation and apply it as a loss term for the autoencoder. The proposed method "RTD-AE" better preserves the global structure and topology of the data manifold than state-of-the-art competitors as measured by linear correlation, triplet distance ranking accuracy, and Wasserstein distance between persistence barcodes.
****
Interpreting Class Conditional GANs with Channel Awareness
Yingqing He,Zhiyi Zhang,Jiapeng Zhu,Yujun Shen,Qifeng Chen
Understanding the mechanism of generative adversarial networks (GANs) helps us better use GANs for downstream applications. Existing efforts mainly target interpreting unconditional models, leaving it less explored how a conditional GAN learns to render images regarding various categories. This work fills in this gap by investigating how a class conditional generator unifies the synthesis of multiple classes. For this purpose, we dive into the widely used class-conditional batch normalization (CCBN), and observe that each feature channel is activated at varying degrees given different categorical embeddings. To describe such a phenomenon, we propose channel awareness, which quantitatively characterizes how a single channel contributes to the final synthesis. Extensive evaluations and analyses on the BigGAN model pre-trained on ImageNet reveal that only a subset of channels is primarily responsible for the generation of a particular category, similar categories (e.g., cat and dog) usually get related to some same channels, and some channels turn out to share information across all classes. For good measure, our algorithm enables several novel applications with conditional GANs. Concretely, we achieve (1) versatile image editing via simply altering a single channel and manage to (2) harmoniously hybridize two different classes. We further verify that the proposed channel awareness shows promising potential in (3) segmenting the synthesized image and (4) evaluating the category-wise synthesis performance.
****
Escaping saddle points in zeroth-order optimization: two function evaluations suffice
Zhaolin Ren,Yujie Tang,Na Li
Zeroth-order methods are useful in solving black-box optimization and reinforcement learning problems in unknown environments. It uses function values to estimate the gradient. As optimization problems are often nonconvex, it is a natural question to understand how zeroth-order methods escape saddle points. In this paper, we consider zeroth-order methods, that at each iteration, may freely choose 2$m$ function evaluations where $m$ ranges from 1 to $d$, with $d$ denoting the problem dimension. We show that by adding an appropriate isotropic perturbation at each iteration, a zeroth-order algorithm based on $2m$ function evaluations per iteration can not only find $\epsilon$-second order stationary points polynomially fast, but do so using only $\tilde{O}(\frac{d}{\epsilon^{2.5}})$ function evaluations.
****
Vector Quantization and Shifting: Exploiting Latent Properties to Optimize Neural Codecs
Muhammet Balcilar,Bharath Bhushan Damodaran,Karam Naser,Franck Galpin,Pierre Hellier
End-to-end image/video codecs are getting competitive compared to traditional compression techniques that have been developed through decades of manual engineering efforts. These trainable codecs have many advantages over traditional techniques such as easy adaptation on perceptual distortion metrics and high performance on specific domains thanks to their learning ability. However, state of the art neural codecs do not take advantage of vector quantization technique and existence of gradient of entropy in decoding device. In this research, we propose some theoretical insights about these two properties (quantization and entropy gradient), and show that this can improve the performances of many off-the-shelf codecs. First, we prove that non-uniform quantization map on neural codec's latent

is not necessary. Thus, we improve the performance by using a predefined optimal uniform vector quantization map. Secondly, we theoretically show that gradient of entropy (available at decoder side) is correlated with the gradient of the reconstruction error (which is not available at decoder side). Thus, we use the former as a proxy in order to improve the compression performance. According to our results, we save between 2-4\% of rate for the same quality with this proposal, for various pre-trained methods.

****

Time-Myopic Go-Explore: Learning A State Representation for the Go-Explore Paradigm
Marc Höftmann,Jan Robine,Stefan Harmeling
Very large state spaces with a sparse reward signal are difficult to explore. The lack of a sophisticated guidance results in a poor performance for numerous reinforcement learning algorithms. In these cases, the commonly used random exploration is often not helpful. The literature shows that this kind of environments require enormous efforts to systematically explore large chunks of the state space. Learned state representations can help here to improve the search by providing semantic context and build a structure on top of the raw observations. In this work we introduce a novel time-myopic state representation that clusters temporal close states together while providing a time prediction capability between them. By adapting this model to the Go-Explore paradigm (Ecoffet et al., 2021b), we demonstrate the first learned state representation that reliably estimates novelty instead of using the hand-crafted representation heuristic. Our method shows an improved solution for the detachment problem which still remains an issue at the Go-Explore Exploration Phase. We provide evidence that our proposed method covers the entire state space with respect to all possible time trajectories — without causing disadvantageous conflict-overlaps in the cell archive. Analogous to native Go-Explore, our approach is evaluated on the hard exploration environments MontezumaRevenge, Gravitar and Frostbite (Atari) in order to validate its capabilities on difficult tasks. Our experiments show that time-myopic Go-Explore is an effective alternative for the domain-engineered heuristic while also being more general. The source code of the method is available on GitHub.

****

Mastering Spatial Graph Prediction of Road Networks
Sotiris Anagnostidis,Aurelien Lucchi,Thomas Hofmann
Accurately predicting road networks from satellite images requires a global understanding of the network topology. We propose to capture such high-level information by introducing a graph-based framework that simulates the addition of sequences of graph edges using a reinforcement learning (RL) approach. In particular, given a partially generated graph associated with a satellite image, an RL agent nominates modifications that maximize a cumulative reward. As opposed to standard supervised techniques that tend to be more restricted to commonly used surrogate losses, these rewards can be based on various complex, potentially non-continuous, metrics of interest. This yields more power and flexibility to encode problem-dependent knowledge. Empirical results on several benchmark datasets demonstrate enhanced performance and increased high-level reasoning about the graph topology when using a tree-based search. We further highlight the superiority of our approach under substantial occlusions by introducing a new synthetic benchmark dataset for this task.

****

Learning Probabilistic Topological Representations Using Discrete Morse Theory
Xiaoling Hu,Dimitris Samaras,Chao Chen
Accurate delineation of fine-scale structures is a very important yet challenging problem. Existing methods use topological information as an additional training loss, but are ultimately making pixel-wise predictions. In this paper, we propose a novel deep learning based method to learn topological/structural. We use discrete Morse theory and persistent homology to construct a one-parameter family of structures as the topological/structural representation space. Furthermore, we learn a probabilistic model that can perform inference tasks in such a topological/structural representation space. Our method generates true structures rath

er than pixel-maps, leading to better topological integrity in automatic segmentation tasks. It also facilitates semi-automatic interactive annotation/proofreading via the sampling of structures and structure-aware uncertainty.
****

Zero-Label Prompt Selection
Chonghua Liao,Yanan Zheng,Zhilin Yang
Natural language prompts have been shown to facilitate cross-task generalization for large language models. However, with no or limited labeled examples, the cross-task performance is highly sensitive to the choice of prompts, while selecting a high-performing prompt is challenging given the scarcity of labels. To address the issue, we propose a Zero-Label Prompt Selection (ZPS) method that selects prompts without any labeled data or gradient update. Specifically, given the candidate human-written prompts for a task, ZPS labels a set of unlabeled data with a prompt ensemble and uses the pseudo-labels for prompt selection. Experiments show that ZPS improves over prior methods by a sizeable margin in zero-label performance. We also extend ZPS to a few-shot setting and show its advantages over strong baselines such as prompt tuning and model tuning.
****

The Curious Case of Benign Memorization
Sotiris Anagnostidis,Gregor Bachmann,Lorenzo Noci,Thomas Hofmann
Despite the empirical advances of deep learning across a variety of learning tasks, our theoretical understanding of its success is still very restricted. One of the key challenges is the overparametrized nature of modern models, enabling complete overfitting of the data even if the labels are randomized, i.e. networks can completely \textit{memorize} all given patterns. While such a memorization capacity seems worrisome, in this work we show that under training protocols that include \textit{data augmentation}, neural networks learn to memorize entirely random labels in a benign way, i.e. they learn embeddings that lead to highly non-trivial performance under nearest neighbour probing. We demonstrate that deep models have the surprising ability to separate noise from signal by distributing the task of memorization and feature learning to different layers. As a result, only the very last layers are used for memorization, while preceding layers encode performant features which remain largely unaffected by the label noise. We explore the intricate role of the augmentations used for training and identify a memorization-generalization trade-off in terms of their diversity, marking a clear distinction to all previous works. Finally, we give a first explanation for the emergence of benign memorization by showing that \textit{malign} memorization under data augmentation is infeasible due to the insufficient capacity of the model for the increased sample size. As a consequence, the network is forced to leverage the correlated nature of the augmentations and as a result learns meaningful features. To complete the picture, a better theory of feature learning in deep neural networks is required to fully understand the origins of this phenomenon.
****

A Connection between One-Step Regularization and Critic Regularization in Reinforcement Learning
Benjamin Eysenbach,Matthieu Geist,Sergey Levine,Ruslan Salakhutdinov
As with any machine learning problem with limited data, effective offline RL algorithms require careful regularization to avoid overfitting. One-step methods perform regularization by doing just a single step of policy improvement, while critic regularization methods do many steps of policy improvement with a regularized objective. These methods appear distinct. One-step methods, such as advantage-weighted regression and conditional behavioral cloning, truncate policy iteration after just one step. This ``early stopping'' makes one-step RL simple and stable, but can limit its asymptotic performance. Critic regularization typically requires more compute but has appealing lower-bound guarantees. In this paper, we draw a close connection between these methods: applying a multi-step critic regularization method with a regularization coefficient of 1 yields the same policy as one-step RL. While practical implementations violate our assumptions and critic regularization is typically applied with smaller regularization coefficients

, our experiments nevertheless show that our analysis makes accurate, testable predictions about practical offline RL methods (CQL and one-step RL) with commonly-used hyperparameters. Our results that every problem can be solved with a single step of policy improvement, but rather that one-step RL might be competitive with critic regularization on RL problems that demand strong regularization.

****

Deep Class Conditional Gaussians for Continual Learning

Thomas L Lee,Amos Storkey

The current state of the art for continual learning with frozen, pre-trained embedding networks are simple probabilistic models defined over the embedding space, for example class conditional Gaussians. However, as of yet, in the task-incremental online setting, it has been an open question how to extend these methods to when the embedding function has to be learned from scratch. In this paper, we propose an empirical Bayesian framework that works by storing a fixed number of examples in memory which are used to calculate the posterior of the probabilistic model and a conditional marginal likelihood term used to fit the embedding function. The learning of the embedding function can be interpreted as using a variant of experience replay, which is a highly performative method for continual learning. As part of our framework, we decide which examples to store by selecting the subset that minimises the KL divergence between the true posterior and the posterior induced by the subset, which is shown to be necessary to achieve good performance. We demonstrate the performance of our method on a range of task-incremental online settings, including those with overlapping tasks which thus far have been under-explored. Our method outperforms all other methods, including several other replay-based methods, evidencing the potential of our approach.

****

Unbiased Supervised Contrastive Learning

Carlo Alberto Barbano,Benoit Dufumier,Enzo Tartaglione,Marco Grangetto,Pietro Gori

Many datasets are biased, namely they contain easy-to-learn features that are highly correlated with the target class only in the dataset but not in the true underlying distribution of the data. For this reason, learning unbiased models from biased data has become a very relevant research topic in the last years. In this work, we tackle the problem of learning representations that are robust to biases. We first present a margin-based theoretical framework that allows us to clarify why recent contrastive losses (InfoNCE, SupCon, etc.) can fail when dealing with biased data. Based on that, we derive a novel formulation of the supervised contrastive loss (ϵ-SupInfoNCE), providing more accurate control of the minimal distance between positive and negative samples.
Furthermore, thanks to our theoretical framework, we also propose FairKL, a new debiasing regularization loss, that works well even with extremely biased data. We validate the proposed losses on standard vision datasets including CIFAR10, CIFAR100, and ImageNet, and we assess the debiasing capability of FairKL with ϵ-SupInfoNCE, reaching state-of-the-art performance on a number of biased datasets, including real instances of biases "in the wild".

****

ReaKE: Contrastive Molecular Representation Learning with Chemical Synthetic Knowledge Graph

Yi Wang,Shuangjia Zheng,Jiahua Rao,Yunan Luo,Yuedong Yang

Molecular representation learning has demonstrated great promise in bridging machine learning and chemical science and in supporting novel chemical discoveries. State-of-the-art methods mostly employ graph neural networks (GNNs) with self-supervised learning (SSL) and extra chemical reaction knowledge to empower the learned embeddings. However, prior works ignore three major issues in modeling reaction data, that is abnormal energy flow, ambiguous embeddings, and sparse embedding space problems. To address these problems, we propose ReaKE, a chemical synthetic knowledge graph-driven pre-training framework for molecular representation learning. We first construct a large-scale chemical synthetic knowledge graph comprising reactants, products and reaction rules. We then propose triplet-level and graph-level contrastive learning strategies to jointly optimize the knowled ge graph and molecular embeddings. Representations learned by ReaKE can capture intermolecular relationships reflected in the semantic knowledge graph and molecular structures. By comparing with other state-of-the-art methods, we show that ReaKE can achieve competitive performance on the reaction prediction pretext task and the learned representations transfer well to various downstream tasks, including reaction classification, yield prediction, and molecule property prediction. Further visualization shows that the learned representations can capture the fine-grained differences both between reactions and between molecules.

****

Graph MLP-Mixer

Xiaoxin He,Bryan Hooi,Thomas Laurent,Adam Perold,Yann LeCun,Xavier Bresson

Graph Neural Networks (GNNs) have shown great potential in the field of graph representation learning. Standard GNNs define a local message-passing mechanism which propagates information over the whole graph domain by stacking multiple layers. This paradigm suffers from two major limitations, over-squashing and poor long-range dependencies, that can be solved using global attention but significantly increases the computational cost to quadratic complexity. In this work, we consider an alternative approach to overcome these structural limitations while keeping a low complexity cost. Motivated by the recent MLP-Mixer architecture introduced in computer vision, we propose to generalize this network to graphs. This GNN model, namely Graph MLP-Mixer, can make long-range connections without over-squashing or high complexity due to the mixer layer applied to the graph patches extracted from the original graph. As a result, this architecture exhibits promising results when comparing standard GNNs vs. Graph MLP-Mixers on benchmark graph datasets.

****

Learning to Register Unbalanced Point Pairs

Kanghee Lee,Junha Lee,Jaesik Park

Point cloud registration methods can effectively handle large-scale, partially overlapping point cloud pairs. Despite its practicality, matching the unbalanced pairs in terms of spatial extent and density has been overlooked and rarely studied. We present a novel method, dubbed UPPNet, for Unbalanced Point cloud Pair registration. We propose to incorporate a hierarchical framework that effectively finds inlier correspondences by gradually reducing search space. The proposed method first predicts subregions within target point cloud that are likely to be overlapped with query. Then following super-point matching and fine-grained refinement modules predict accurate inlier correspondences between the target and query. Additional geometric constraints are applied to refine the correspondences that satisfy spatial compatibility. The proposed network can be trained in an end-to-end manner, predicting the accurate rigid transformation with a single forward pass. To validate the efficacy of the proposed method, we create a carefully designed benchmark, named KITTI-UPP dataset, by augmenting the KITTI odometry dataset. Extensive experiments reveal that the proposed method not only outperforms state-of-the-art point cloud registration methods by large margins on KITTI-UPP benchmark, but also achieves competitive results on the standard pairwise registration benchmark including 3DMatch, 3DLoMatch, ScanNet, and KITTI, thus showing the applicability of our method on various datasets. The source code and dataset will be publicly released.

****

On Feature Diversity in Energy-based Models

Firas Laakom,Jenni Raitoharju,Alexandros Iosifidis,Moncef Gabbouj

Energy-based learning is a powerful learning paradigm that encapsulates various discriminative and generative approaches. An energy-based model (EBM) is typically formed of inner-model(s) that learn a combination of the different features to generate an energy mapping for each input configuration. In this paper, we focus on the diversity of the produced feature set. We extend the probably approximately correct (PAC) theory of EBMs and analyze the effect of redundancy reduction on the performance of EBMs. We derive generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and we show that indeed reducing redundancy of the featu re set can consistently decrease the gap between the true and empirical expectation of the energy and boosts the performance of the model.

****

Physics Model-based Autoencoding for Magnetic Resonance Fingerprinting

Juyeon Heo,Pingfan Song,Weiyang Liu,Adrian Weller

Magnetic Resonance Fingerprinting (MRF) is a promising paradigm to achieve fast quantitative Magnetic Resonance Imaging (QMRI). However, current MRF methods suffer from slow imaging speeds and poor generalization performance on radio frequency pulse sequences generated with varied settings. To address this challenging task, we propose a novel model-based MRF method that learns better representations by integrating a fast and differentiable MRI physics model as causal regularization. The proposed approach adopts a supervised auto-encoder framework consisting of an encoder and a decoder, where the encoder predicts the target tissue properties (anti-causal task) and the decoder reconstructs the inputs (causal task). Specifically, the encoder embeds high-dimensional MRF time sequences to a low-dimensional tissue property space, while the decoder exploits an MRI physics model to reconstruct the input signals using the estimated tissue properties and associated MRI settings. The causal regularization induced by the decoder improves the generalization performance and uniform stability of the approach, leading to the best performance on tissue property estimation, outperforming state-of-the-art competing methods.

****

Compositional Prompt Tuning with Motion Cues for Open-vocabulary Video Relation Detection

Kaifeng Gao,Long Chen,Hanwang Zhang,Jun Xiao,Qianru Sun

Prompt tuning with large-scale pretrained vision-language models empowers open-vocabulary prediction trained on limited base categories, e.g., object classification and detection. In this paper, we propose compositional prompt tuning with motion cues: an extended prompt tuning paradigm for compositional predictions of video data. In particular, we present Relation Prompt (RePro) for Open-vocabulary Video Visual Relation Detection (Open-VidVRD), where conventional prompt tuning is easily biased to certain subject-object combinations and motion patterns. To this end, RePro addresses the two technical challenges of Open-VidVRD: 1) the prompt tokens should respect the two different semantic roles of subject and object, and 2) the tuning should account for the diverse spatiotemporal motion patterns of the subject-object compositions. Our RePro achieves a new state-of-the-art performance on two VidVRD benchmarks of not only the base training object and predicate categories, but also the unseen ones. Extensive ablations also demonstrate the effectiveness of the proposed compositional and multi-mode design of prompt. Code is available at https://github.com/Dawn-LX/OpenVoc-VidVRD.

****

Multi-objective optimization via equivariant deep hypervolume approximation

Jim Boelrijk,Bernd Ensing,Patrick Forré

Optimizing multiple competing objectives is a common problem across science and industry. The inherent inextricable trade-off between those objectives leads one to the task of exploring their Pareto front. A meaningful quantity for the purpose of the latter is the hypervolume indicator, which is used in Bayesian Optimization (BO) and Evolutionary Algorithms (EAs). However, the computational complexity for the calculation of the hypervolume scales unfavorably with increasing number of objectives and data points, which restricts its use in those common multi-objective optimization frameworks.

To overcome these restrictions, previous work has focused on approximating the hypervolume using deep learning. In this work, we propose a novel deep learning architecture to approximate the hypervolume function, which we call DeepHV. For better sample efficiency and generalization, we exploit the fact that the hypervolume is scale equivariant in each of the objectives as well as permutation invariant w.r.t. both the objectives and the samples, by using a deep neural network that is equivariant w.r.t. the combined group of scalings and permutations. We show through an ablation study that including these symmetries leads to significantly improved model accuracy.

We evaluate our method against exact, and approximate hypervolume methods in terms of accuracy, computation time, and generalization. We also apply and compare our methods to state-of-the-art multi-objective BO methods and EAs on a range of synthetic and real-world benchmark test cases. The results show that our methods are promising for such multi-objective optimization tasks.

****

Conditional Execution Of Cascaded Models Improves The Accuracy-Efficiency Trade-Off

Luzian Lebovitz,Lukas Cavigelli,Michele Magno,Lorenz K Muller

The compute effort required to perform inference on state-of-the-art deep learning models is ever growing. Practical applications are commonly limited to a certain cost per inference. Cascades of pretrained models with conditional execution address these requirements based on the intuition that some inputs are easy enough that they can be processed correctly by a small model allowing for an early exit. If the small model is not sufficiently confident in its prediction, the input is passed on to a larger model. The selection of the confidence threshold allows to trade off compute effort against accuracy. In this work, we explore the effective design of model cascades, and thoroughly evaluate the impact on the accuracy-compute trade-off. We find that they not only interpolate favorably between pretrained models, but that this trade-off curve commonly outperforms single models. This allows us to redefine most of the ImageNet Pareto front already with 2-model cascades, achieving an average reduction in compute effort at equal accuracy of almost 3.1x above 86% and more than 1.9x between 80% and 86% top-1 accuracy. We confirm the wide applicability and effectiveness of the method on the GLUE benchmark. We release the code to reproduce our experiments in the supplementary material and use only publicly available models and datasets.

****

Adversarial Text to Continuous Image Generation

Kilichbek Haydarov,Aashiq Muhamed,Jovana Lazarevic,Ivan Skorokhodov,Xiaoqian Shen,Chamuditha Jayanga Galappaththige,Mohamed Elhoseiny

Implicit Neural Representations (INR) provide a natural way to parametrize images as a continuous signal, using an MLP that predicts the RGB color at an (x, y) image location. Recently, it has been demonstrated that high-quality INR-decoders can be designed and integrated with Generative Adversarial Networks (GANs) to facilitate unconditional continuous image generation, that are no longer bounded to a spatial resolution. In this paper, we introduce HyperCGAN, a conceptually simple approach for Adversarial Text to Continuous Image Generation based on HyperNetworks, which are networks that produce parameters for another network. HyperCGAN utilizes HyperNetworks to condition an INR-based GAN model on text. In this setting, the generator and the discriminator weights are controlled by their corresponding HyperNetworks, which modulate weight parameters using the provided text query. We propose an effective Word-level hyper-modulation Attention operator, termed WhAtt, which encourages grounding words to independent pixels at input (x, y) coordinates. To the best of our knowledge, our work is the first that explores text-controllable continuous image generation. We conduct comprehensive experiments on the COCO 256x256, CUB 256x256, and the ArtEmis 256x256 benchmark which we introduce in this paper. HyperCGAN improves the performance of text-controllable image generators over the baselines while significantly reducing the gap between text-to-continuous and text-to-discrete image synthesis. Additionally, we show that HyperCGAN, when conditioned on text, retains the desired properties of continuous generative models (e.g., extrapolation outside of image boundaries, accelerated inference of low-resolution images, out-of-the-box superresolution).

****

Fine-grained Few-shot Recognition by Deep Object Parsing

Ruizhao Zhu,Pengkai Zhu,Samarth Mishra,Venkatesh Saligrama

We propose a new method for fine-grained few-shot recognition via deep object parsing. In our framework, an object is made up of K distinct parts and for each part, we learn a dictionary of templates, which is shared across all instances and categories. An object is parsed by estimating the locations of these K par ts and a set of active templates that can reconstruct the part features. We recognize test instances by comparing its active templates and the relative geometry of its part locations against those of the presented few-shot instances. Our method is end-to-end trainable to learn part templates on-top of a convolutional backbone. To combat visual distortions such as orientation, pose and size, we learn templates at multiple scales, and at test-time parse and match instances across these scales. We show that our method is competitive with the state-of-the-art, and by virtue of parsing enjoys interpretability as well.

****

Can Wikipedia Help Offline Reinforcement Learning?

Machel Reid,Yutaro Yamada,Shixiang Shane Gu

Fine-tuning reinforcement learning (RL) models has been challenging because of a lack of large scale off-the-shelf datasets as well as high variance in transferability among different environments. Recent work has looked at tackling offline RL from the perspective of sequence modeling with improved results as result of the introduction of the Transformer architecture. However, when the model is trained from scratch, it suffers from slow convergence speeds. In this paper, we look to take advantage of this formulation of reinforcement learning as sequence modeling and investigate the transferability of pre-trained sequence models on other domains (vision, language) when finetuned on offline RL tasks (control, games). To this end, we also propose techniques to improve transfer between these domains. Results show consistent performance gains in terms of both convergence speed and reward on a variety of environments, accelerating training by 3-6x and achieving state-of-the-art performance in a variety of tasks using Wikipedia-pretrained and GPT2 language models. We hope that this work not only brings light to the potentials of leveraging generic sequence modeling techniques and pre-trained models for RL, but also inspires future work on sharing knowledge between generative modeling tasks of completely different domains.

****

Lightweight Equivariant Graph Representation Learning for Protein Engineering

Bingxin Zhou,Outongyi Lv,Kai Yi,Xinye Xiong,Pan Tan,Liang Hong,Yu Guang Wang

This work tackles the issue of directed evolution in computational protein design that makes accurate predictions of the function of a protein mutant. We design a lightweight pre-training graph neural network model for multi-task protein representation learning from its 3D structure. Rather than reconstructing and optimizing the protein structure, the trained model recovers the amino acid types and key properties of the central residues from a given noisy three-dimensional local environment. On the prediction task for the higher-order mutants, where many amino acid sites of the protein are mutated, the proposed training strategy achieves remarkably higher performance by 20% improvement at the cost of requiring less than 1% of computational resources that are required by popular transformer-based state-of-the-art deep learning models for protein design.

****

DiffusER: Diffusion via Edit-based Reconstruction

Machel Reid,Vincent Josua Hellendoorn,Graham Neubig

In text generation, models that generate text from scratch one token at a time are currently the dominant paradigm. Despite being performant, these models lack the ability to revise existing text, which limits their usability in many practical scenarios. We look to address this, with DiffusER (Diffusion via Edit-based Reconstruction), a new edit-based generative model for text based on denoising diffusion models -- a class of models that use a Markov chain of denoising steps to incrementally generate data. DiffusER is not only a strong generative model in general, rivalling autoregressive models on several tasks spanning machine translation, summarization, and style transfer; it can also perform other varieties of generation that standard autoregressive models are not well-suited for. For instance, we demonstrate that DiffusER makes it possible for a user to condition generation on a prototype, or an incomplete sequence, and continue revising based on previous edit steps.

****

Modeling Temporal Data as Continuous Functions with Process Diffusion

Marin Biloš,Kashif Rasul,Anderson Schneider,Yuriy Nevmyvaka,Stephan Günnemann

Temporal data like time series are often observed at irregular intervals which is a challenging setting for the existing machine learning methods. To tackle this problem, we view such data as samples from some underlying continuous function. We then define a diffusion-based generative model that adds noise from a predefined stochastic process while preserving the continuity of the resulting underlying function. A neural network is trained to reverse this process which allows us to sample new realizations from the learned distribution. We define suitable stochastic processes as noise sources and introduce novel denoising and score-matching models on processes. Further, we show how to apply this approach to the multivariate probabilistic forecasting and imputation tasks. Through our extensive experiments, we demonstrate that our method outperforms previous models on synthetic and real-world datasets.

****

KeyCLD: Learning Constrained Lagrangian Dynamics in Keypoint Coordinates from Images

Rembert Daems,Jeroen Taets,Francis wyffels,Guillaume Crevecoeur

We present KeyCLD, a framework to learn Lagrangian dynamics from images. Learned keypoint representations derived from images are directly used as positional state vector for jointly learning constrained Lagrangian dynamics. KeyCLD is trained unsupervised end-to-end on sequences of images. Our method explicitly models the mass matrix, potential energy and the input matrix, thus allowing energy based control. We demonstrate learning of Lagrangian dynamics from images on the dm_control pendulum, cartpole and acrobot environments, wether they are unactuated, underactuated or fully actuated. Trained models are able to produce long-term video predictions, showing that the dynamics are accurately learned. Our method strongly outperforms recent works on learning Lagrangian or Hamiltonian dynamics from images. The benefits of including a Lagrangian prior and prior knowledge of a constraint function is further investigated and empirically evaluated.

****

DynaMS: Dyanmic Margin Selection for Efficient Deep Learning

Jiaxing Wang,Yong Li,Jingwei Zhuo,Xupeng Shi,WEIZHONG ZHANG,Lixing Gong,Tong Tao,Pengzhang Liu,Yongjun Bao,Weipeng Yan

The great success of deep learning is largely driven by training over-parameterized models on massive datasets. To avoid excessive computation, extracting and training only on the most informative subset is drawing increasing attention. Nevertheless, it is still an open question how to select such a subset on which the model trained generalizes on par with the full data. In this paper, we propose dynamic margin selection (DynaMS). DynaMS leverages the distance from candidate samples to the classification boundary to construct the subset, and the subset is dynamically updated during model training. We show that DynaMS converges with large probability, and for the first time show both in theory and practice that dynamically updating the subset can result in better generalization over previous works. To reduce the additional computation incurred by the selection, a light parameter sharing proxy (PSP) is designed. PSP is able to faithfully evaluate instances with respect to the current model, which is necessary for dynamic selection. Extensive analysis and experiments demonstrate the superiority of the proposed approach in data selection against many state-of-the-art counterparts on benchmark datasets.

****

TANGOS: Regularizing Tabular Neural Networks through Gradient Orthogonalization and Specialization

Alan Jeffares,Tennison Liu,Jonathan Crabbé,Fergus Imrie,Mihaela van der Schaar

Despite their success with unstructured data, deep neural networks are not yet a panacea for structured tabular data. In the tabular domain, their efficiency crucially relies on various forms of regularization to prevent overfitting and provide strong generalization performance. Existing regularization techniques include broad modelling decisions such as choice of architecture, loss functions, and optimization methods. In this work, we introduce Tabular Neural Gradient Orthogonalization and Specialization (TANGOS), a novel framework for regularization in the tabular setting built on latent unit attributions. The gradient attribution of an activation with respect to a given input feature suggests how the neuron attends to that feature, and is often employed to interpret the predictions of deep networks. In TANGOS, we take a different approach and incorporate neuron attributions directly into training to encourage orthogonalization and specialization of latent attributions in a fully-connected network. Our regularizer encourages neurons to focus on sparse, non-overlapping input features and results in a set of diverse and specialized latent units. In the tabular domain, we demonstrate that our approach can lead to improved out-of-sample generalization performance, outperforming other popular regularization methods. We provide insight into why our regularizer is effective and demonstrate that TANGOS can be applied jointly with existing methods to achieve even greater generalization performance.

****

How does Uncertainty-aware Sample-selection Help Decision against Action Noise?

Xingrui Yu,Bo Han,Ivor Tsang

Learning from imperfect demonstrations has become a vital problem in imitation learning (IL). Since the assumption of the collected demonstrations are optimal cannot always hold in real-world tasks, many previous works considers learning from a mixture of optimal and sub-optimal demonstrations. On the other hand, video records can be hands-down demonstrations in practice. Leveraging such demonstrations requires labors to output action for each frame. However, action noise always occurs when the labors are not domain experts, or meet confusing state frames. Previous IL methods can be vulnerable to such demonstrations with state-dependent action noise. To tackle this problem, we propose a robust learning paradigm called USN, which bridges Uncertainty-aware Sample-selection with Negative learning. First, IL model feeds forward all demonstration data and estimates its predictive uncertainty. Then, we select large-loss samples in the light of the uncertainty measures. Next, we update the model parameters with additional negative learning on the selected samples. Empirical results on Box2D tasks and Atari games demonstrate that USN improves the performance of state-of-the-art IL methods by more than 10% under a large portion of action noise.

****

Inversely Eliciting Numerical Reasoning in Language Models via Solving Linear Systems

Fan Zhou,Haoyu Dong,Qian Liu,Zhoujun Cheng,Shi Han,Dongmei Zhang

Numerical reasoning over natural language has been a long-standing goal for the research community. However, recent language models have proven difficult to reliably generalize to a broad range of numbers, although they have shown proficiency in reasoning over common and simple numbers. In this paper, we propose a novel method to elicit and exploit the numerical reasoning knowledge hidden in pre-trained language models using simple anchor numbers. Concretely, we first leverage simple numbers as anchors to probe the implicitly inferred arithmetic expressions from language models, and then explicitly apply the expressions on complex numbers to get corresponding answers. To inversely elicit arithmetic expressions, we transform and formulate the task as an analytically solvable linear system. Experimental results on several numerical reasoning benchmarks demonstrate that our approach is highly effective. More importantly, our approach works in the inference phase without extra model training, making it highly portable and achieving significant and consistent performance benefits across a variety of language models in zero-shot, few-shot, and fine-tuning scenarios.

****

Model-based Causal Bayesian Optimization

Scott Sussex,Anastasia Makarova,Andreas Krause

How should we intervene on an unknown structural equation model to maximize a downstream variable of interest? This setting, also known as causal Bayesian optimization (CBO), has important applications in medicine, ecology, and manufacturing. Standard Bayesian optimization algorithms fail to effectively leverage the underlying causal structure. Existing CBO approaches assume noiseless measurements and do not come with guarantees. We propose the {\em model-based causal Bayesian optimization algorithm (MCBO)} that learns a full system model instead of only modeling intervention-reward pairs. MCBO propagates epistemic uncertainty about the causal mechanisms through the graph and trades off exploration and exploitation via the optimism principle. We bound its cumulative regret, and obtain the first non-asymptotic bounds for CBO. Unlike in standard Bayesian optimization, our acquisition function cannot be evaluated in closed form, so we show how the reparameterization trick can be used to apply gradient-based optimizers. The resulting practical implementation of MCBO compares favorably with state-of-the-art approaches empirically.

****

Targeted Attacks on Timeseries Forecasting

Yuvaraj Govindarajulu,Avinash Amballa,Pavan Kulkarni,Manojkumar Parmar

Real-world deep learning models built for Time Series Forecasting are used in several critical applications from medical devices to the security domain. Many previous works have shown how deep learning models are prone to adversarial attacks and studied their vulnerabilities. However, the vulnerabilities of time series models for forecasting due to adversarial inputs are not extensively studied. While attack on a forecasting model might be intended to deteriorate the performance of the model, it is more effective, if the attack is focused on a specific impact on the model's output. In this paper, we propose a novel formulation of Directional, Amplitudinal, and Temporal targeted adversarial attacks on time series forecasting models. These targeted attacks create a specific impact or hide a potential high-impact area on the forecasting output. We use the existing adversarial attack techniques from the computer vision domain and adapt them for time series. Additionally, we propose a modified version of the Auto Projected Gradient Descent attack for targeted attacks. We explore the impact of the proposed targeted attacks against untargeted attacks. We use KS-Tests to statistically prove the impact of the attack. Our experimental results demonstrate how targeted attacks on time series models are practical and are more powerful in terms of statistical similarity. It is, hence difficult to detect through statistical methods. We believe that this work opens a new paradigm in the time series forecasting domain and is an important consideration for developing better defenses.

****

QuAFL: Federated Averaging Made Asynchronous and Communication-Efficient

Hossein Zakerinia,Shayan Talaei,Giorgi Nadiradze,Dan Alistarh

Federated Learning (FL) is an emerging paradigm to enable the large-scale distributed training of machine learning models, while still providing privacy guarantees.

In this work, we address two of the main practical challenges when scaling federated optimization to large node counts: the need for tight synchronization between the central authority and individual computing nodes, and the large communication cost of transmissions between the central server and clients.

Specifically, we present a new variant of the classic federated averaging (FedAvg) algorithm, which supports both asynchronous communication and communication compression. We provide a new analysis technique showing that, in spite of these system relaxations, our algorithm can provide similar convergence to FedAvg in some parameter regimes.

On the experimental side, we show that our algorithm ensures fast convergence for standard federated tasks.

****

Random Matrix Analysis to Balance between Supervised and Unsupervised Learning under the Low Density Separation Assumption

Vasilii Feofanov,Malik Tiomoko,Aladin Virmaux

We propose a theoretical framework to analyze semi-supervised classification under the low density separation assumption in a high-dimensional regime. In particular, we introduce QLDS, a linear classification model, where the low density separation assumption is implemented via quadratic margin maximization.

The algorithm has an explicit solution with rich theoretical properties, and we show that particular cases of our algorithm are the least-square support vector machine in the supervised case, the spectral clustering in the fully unsupervised regime, and a class of semi-supervised graph-based approaches. As such, QLDS e stablishes a smooth bridge between these supervised and unsupervised learning methods. Using recent advances in the random matrix theory, we formally derive a theoretical evaluation of the classification error in the asymptotic regime.

As an application, we derive a hyperparameter selection policy that finds the best balance between the supervised and the unsupervised terms of our learning criterion.

Finally, we provide extensive illustrations of our framework, as well as an experimental study on several benchmarks to demonstrate that QLDS, while being computationally more efficient, improves over cross-validation for hyperparameter selection, indicating a high promise of the usage of random matrix theory for semi-supervised model selection.

****

Learning to Solve Constraint Satisfaction Problems with Recurrent Transformer

Zhun Yang,Adam Ishay,Joohyung Lee

Constraint satisfaction problems (CSPs) are about finding values of variables that satisfy the given constraints. We show that Transformer extended with recurrence is a viable approach to learning to solve CSPs in an end-to-end manner, having clear advantages over state-of-the-art methods such as Graph Neural Networks, SATNet, and some neuro-symbolic models. With the ability of Transformer to handle visual input, the proposed Recurrent Transformer can straightforwardly be applied to visual constraint reasoning problems while successfully addressing the symbol grounding problem. We also show how to leverage deductive knowledge of discrete constraints in the Transformer's inductive learning to achieve sample-efficient learning and semi-supervised learning for CSPs.

****

MARLlib: Extending RLlib for Multi-agent Reinforcement Learning

Siyi Hu,Yifan Zhong,Minquan Gao,Weixun Wang,Hao Dong,Zhihui Li,Xiaodan Liang,Yaodong Yang,Xiaojun Chang

Despite the fast development of multi-agent reinforcement learning (MARL) methods, there is a lack of commonly-acknowledged baseline implementation and evaluation platforms. As a result, an urgent need for MARL researchers is to develop an integrated library suite, similar to the role of RLlib in single-agent RL, that delivers reliable MARL implementation and replicable evaluation in various bechmarks. To fill such a research gap, in this paper, we propose Multi-Agent RLlib (MARLlib), a comprehensive MARL algorithm library that facilitates RLlib for solving multi-agent problems. With a novel design of agent-level distributed dataflow, MARLlib manages to unify tens of algorithms, including different types of independent learning, centralized critic, and value decomposition methods; this leads to a highly composable integration of MARL algorithms that are not possible to unify before. Furthermore, MARLlib goes beyond current work by integrating diverse environment interfaces and providing flexible parameter sharing strategies; this allows to create versatile solutions to cooperative, competitive, and mixed tasks with minimal code modifications for end users. A plethora of experiments are conducted to substantiate the correctness of our implementation, based on which we further derive new insights on the relationship between the performance and the design of algorithmic components. With MARLlib, we expect researchers to be able to tackle broader real-world multi-agent problems with trustworthy solutions. Our code\footnote{\url{https://github.com/ICLR2023Paper4242/MARLlib}} and documentation\footnote{\url{https://iclr2023marllib.readthedocs.io/}} are released for reference.

****

Improving the imputation of missing data with Markov Blanket discovery

Yang Liu,Anthony Constantinou

The process of imputation of missing data typically relies on generative and regression models. These approaches often operate on the unrealistic assumption that all of the data features are directly related with one another, and use all of the available features to impute missing values. In this paper, we propose a novel Markov Blanket discovery approach to determine the optimal feature set for a given variable by considering both observed variables and missingness of partially observed variables to account for systematic missingness. We then incorporat e this method to the learning process of the state-of-the-art MissForest imputation algorithm, such that it informs MissForest which features to consider to impute missing values, depending on the variable the missing value belongs to. Experiments across different case studies and multiple imputation algorithms show that the proposed solution improves imputation accuracy, both under random and systematic missingness.

****

Boosting the Cycle Counting Power of Graph Neural Networks with I^2-GNNs
Yinan Huang,Xingang Peng,Jianzhu Ma,Muhan Zhang

Message Passing Neural Networks (MPNNs) are a widely used class of Graph Neural Networks (GNNs). The limited representational power of MPNNs inspires the study of provably powerful GNN architectures. However, knowing one model is more powerful than another gives little insight about what functions they can or cannot express. It is still unclear whether these models are able to approximate specific functions such as counting certain graph substructures, which is essential for applications in biology, chemistry and social network analysis. Motivated by this, we propose to study the counting power of Subgraph MPNNs, a recent and popular class of powerful GNN models that extract rooted subgraphs for each node, assign the root node a unique identifier and encode the root node's representation within its rooted subgraph. Specifically, we prove that Subgraph MPNNs fail to count more-than-4-cycles at node level, implying that node representations cannot correctly encode the surrounding substructures like ring systems with more than four atoms. To overcome this limitation, we propose I^2-GNNs to extend Subgraph MPNNs by assigning different identifiers for the root node and its neighbors in each subgraph. I^2-GNNs' discriminative power is shown to be strictly stronger than Subgraph MPNNs and partially stronger than the 3-WL test. More importantly, I^2-GNNs are proven capable of counting all 3, 4, 5 and 6-cycles, covering common substructures like benzene rings in organic chemistry, while still keeping linear complexity. To the best of our knowledge, it is the first linear-time GNN model that can count 6-cycles with theoretical guarantees. We validate its counting power in cycle counting tasks and demonstrate its competitive performance in molecular prediction benchmarks.

****

Energy Consumption-Aware Tabular Benchmarks for Neural Architecture Search
Pedram Bakhtiarifard,Christian Igel,Raghavendra Selvan

The demand for large-scale computational resources for Neural Architecture Search (NAS) has been lessened by tabular benchmarks for NAS. Evaluating NAS strategies is now possible on extensive search spaces and at a moderate computational cost. But so far, NAS has mainly focused on maximising performance on some hold-out validation/test set. However, energy consumption is a partially conflicting objective that should not be neglected. We hypothesise that constraining NAS to include the energy consumption of training the models could reveal a sub-space of undiscovered architectures that are more computationally efficient with a smaller carbon footprint. To support the hypothesis, an existing tabular benchmark for NAS is augmented with the energy consumption of each architecture. We then perform multi-objective optimisation that includes energy consumption as an additional objective. We demonstrate the usefulness of multi-objective NAS for uncovering the trade-off between performance and energy consumption as well as for finding more energy-efficient architectures. The updated tabular benchmark is open-sourced to encourage the further exploration of energy consumption-aware NAS.

****

Fundamental Limits in Formal Verification of Message-Passing Neural Networks
Marco Sälzer,Martin Lange

Output reachability and adversarial robustness are among the most relevant safety properties of neural networks.
We show that in the context of Message Passing Neural Networks (MPNN), a common Graph Neural Network (GNN) model,
formal verification is impossible. In particular, we show that output reachability of graph-classifier MPNN,
working over graphs of unbounded size, non-trivial degree and sufficiently expre

ssive node labels, cannot be verified formally: there is no algorithm that answers correctly (with yes or no), given an MPNN, whether there exists some valid input to the MPNN such that the corresponding output satisfies a given specification. However, we also show that output reachability and adversarial robustness of node-classifier MPNN can be verified formally when a limit on the degree of input graphs is given a priori. We discuss the implications of these results, for the purpose of obtaining a complete picture of the principle possibility to formally verify GNN, depending on the expressiveness of the involved GNN models and input-output specifications.

****

Score Matching via Differentiable Physics

Benjamin Holzschuh,Simona Vegetti,Nils Thuerey

Diffusion models based on stochastic differential equations (SDEs) gradually perturb a data distribution $p(\mathbf{x})$ over time by adding noise to it. A neural network is trained to approximate the score $\nabla_\mathbf{x} \log p_t(\mathbf{x})$ at time t, which can be used to reverse the corruption process. In this paper, we focus on learning the score field that is associated with the time evolution according to a physics operator in the presence of natural non-deterministic physical processes like diffusion. A decisive difference to previous methods is that the SDE underlying our approach transforms the state of a physical system to another state at a later time. For that purpose, we replace the drift of the underlying SDE formulation with a differentiable simulator or a neural network approximation of the physics. At the core of our method, we optimize the so-called probability flow ODE to fit a training set of simulation trajectories inside an ODE solver and solve the reverse-time SDE for inference to sample plausible trajectories that evolve towards a given end state. We demonstrate the competitiveness of our approach for different challenging inverse problems.

****

QUIC-FL: : Quick Unbiased Compression for Federated Learning

Ran Ben-Basat,Shay Vargaftik,Amit Portnoy,Gil Einziger,Yaniv Ben-Itzhak,Michael Mitzenmacher

Distributed Mean Estimation (DME) is a fundamental building block in communication efficient federated learning. In DME, clients communicate their lossily compressed gradients to the parameter server, which estimates the average and updates the model.

State of the art DME techniques apply either unbiased quantization methods, resulting in large estimation errors, or biased quantization methods, where unbiasing the result requires that the server decodes each gradient individually, which markedly slows the aggregation time.

In this paper, we propose QUIC-FL, a DME algorithm that achieves the best of all worlds. QUIC-FL is unbiased, offers fast aggregation time, and is competitive with the most accurate (slow aggregation) DME techniques. To achieve this, we formalize the problem in a novel way that allows us to use standard solvers to design near-optimal unbiased quantization schemes.

****

FedMEKT: Split Multimodal Embedding Knowledge Transfer in Federated Learning

Huy Quang Le,Minh N. H. Nguyen,Choong Seon Hong

Federated Learning (FL) enables a decentralized machine-learning paradigm to collaboratively train a generalized global model without sharing users' private data. However, most existing FL approaches solely utilize single-modal data, thus limiting the systems for exploiting valuable multimodal data in future personalized applications. Furthermore, most FL methods still rely on the labeled data at the client side, which is limited in real-world applications due to the inability of data self-annotation from users. To leverage the representations from different modalities in FL, we propose a novel multimodal FL framework with a semi-supervised learning setting. Specifically, we develop the split multimodal embedding knowledge transfer mechanism in federated learning, namely, FedMEKT, which en ables the personalized and generalized multimodal representations exchange between server and clients using a small multimodal proxy dataset. Hence, FedMEKT iteratively updates the generalized encoders from the collaborative embedding knowledge of each client, such as modality-averaging representations. Thereby, a generalized encoder could guide personalized encoders to enhance the generalization abilities of client models; afterward, personalized classifiers could be trained using the proxy labeled data to perform supervised tasks. Through the extensive experiments on three multimodal human activity recognition tasks, we demonstrate that FedMEKT achieves superior performance in both local and global encoder models on linear evaluation and guarantees user privacy for personal data and model parameters.

****

Short-Term Memory Convolutions

Grzegorz Stefa■ski,Krzysztof Arendt,Pawe■ Daniluk,Bart■omiej Jasik,Artur Szumaczuk

The real-time processing of time series signals is a critical issue for many real-life applications. The idea of real-time processing is especially important in audio domain as the human perception of sound is sensitive to any kind of disturbance in perceived signals, especially the lag between auditory and visual modalities. The rise of deep learning (DL) models complicated the landscape of signal processing. Although they often have superior quality compared to standard DSP methods, this advantage is diminished by higher latency. In this work we propose novel method for minimization of inference time latency and memory consumption, called Short-Term Memory Convolution (STMC) and its transposed counterpart. The main advantage of STMC is the low latency comparable to long short-term memory (LSTM) networks. Furthermore, the training of STMC-based models is faster and more stable as the method is based solely on convolutional neural networks (CNNs). In this study we demonstrate an application of this solution to a U-Net model for a speech separation task and GhostNet model in acoustic scene classification (ASC) task. In case of speech separation we achieved a 5-fold reduction in inference time and a 2-fold reduction in latency without affecting the output quality. The inference time for ASC task was up to 4 times faster while preserving the original accuracy.

****

End-to-End Speech Synthesis Based on Deep Conditional Schrödinger Bridges

Shoule Wu,Ziqiang Shi

Speech synthesis plays an important role in human-computer interaction. Existing methods mainly employ traditional two-stage pipeline, e.g. text-to-speech and vocoder. In this paper, we propose a system called Schr\"on, which can generate speech waves in an end-to-end mamaner by solving Schr\"odinger bridge problems (SBP). In order to make SBP suitable for speech synthesis, we generalize SBP from two aspects. The first generalization makes it possible to accept condition variables, which are used to control the generated speech, and the second generalization allows it to handle variable-size input. Besides these two generalizations, we propose two techniques to fill the large information gap between text and speech waveforms for generating high-quality voice. The first technique is to use a text-mel joint representation as the conditional input of the conditional SBP. The second one is to use a branch network for the generation of mel scores as a regularization, so that the text features will not be degenerated. Experimental results show that Schr\"on achieves state-of-the-art MOS of 4.52 on public data set LJSpeech. Audio samples are available at https://schron.github.io/.

****

LexMAE: Lexicon-Bottlenecked Pretraining for Large-Scale Retrieval

Tao Shen,Xiubo Geng,Chongyang Tao,Can Xu,Xiaolong Huang,Binxing Jiao,Linjun Yang,Daxin Jiang

In large-scale retrieval, the lexicon-weighting paradigm, learning weighted sparse representations in vocabulary space, has shown promising results with high quality and low latency. Despite it deeply exploiting the lexicon-representing capability of pre-trained language models, a crucial gap remains between language modeling and lexicon-weighting retrieval -- the former preferring certain or low-

entropy words whereas the latter favoring pivot or high-entropy words -- becoming the main barrier to lexicon-weighting performance for large-scale retrieval. To bridge this gap, we propose a brand-new pre-training framework, lexicon-bottlenecked masked autoencoder (LexMAE), to learn importance-aware lexicon representations. Essentially, we present a lexicon-bottlenecked module between a normal language modeling encoder and a weakened decoder, where a continuous bag-of-words bottleneck is constructed to learn a lexicon-importance distribution in an unsupervised fashion. The pre-trained LexMAE is readily transferred to the lexicon-weighting retrieval via fine-tuning. On the ad-hoc retrieval benchmark, MS-Marco, it achieves 42.6% MRR@10 with 45.8 QPS for the passage dataset and 44.4% MRR@100 with 134.8 QPS for the document dataset, by a CPU machine. And LexMAE shows state-of-the-art zero-shot transfer capability on BEIR benchmark with 12 datasets.
****
A GNN-Guided Predict-and-Search Framework for Mixed-Integer Linear Programming
Qingyu Han,Linxin Yang,Qian Chen,Xiang Zhou,Dong Zhang,Akang Wang,Ruoyu Sun,Xiaodong Luo
Mixed-integer linear programming (MILP) is widely employed for modeling combinatorial optimization problems. In practice, similar MILP instances with only coefficient variations are routinely solved, and machine learning (ML) algorithms are capable of capturing common patterns across these MILP instances. In this work, we combine ML with optimization and propose a novel predict-and-search framework for efficiently identifying high-quality feasible solutions. Specifically, we first utilize graph neural networks to predict the marginal probability of each variable, and then search for the best feasible solution within a properly defined ball around the predicted solution. We conduct extensive experiments on public datasets, and computational results demonstrate that our proposed framework achieves 51.1% and 9.9% performance improvements to MILP solvers SCIP and Gurobi on primal gaps, respectively.
****
Parameter Averaging for SGD Stabilizes the Implicit Bias towards Flat Regions
Atsushi Nitanda,Ryuhei Kikuchi,Shugo Maeda
Stochastic gradient descent is a workhorse for training deep neural networks due to its excellent generalization performance. Several studies demonstrated this success is attributed to the implicit bias of the method that prefers a flat minimum and developed new methods based on this perspective. Recently, Izmailov et al. (2018) empirically observed that an averaged stochastic gradient descent with a large step size can bring out the implicit bias more effectively and can converge more stably to a flat minimum than the vanilla stochastic gradient descent. In our work, we theoretically justify this observation by showing that the averaging scheme improves the bias-optimization tradeoff coming from the stochastic gradient noise: a large step size amplifies the bias but makes convergence unstable, and vice versa. Specifically, we show that the averaged stochastic gradient descent can get closer to a solution of a penalized objective on the sharpness than the vanilla stochastic gradient descent using the same step size under certain conditions. In experiments, we verify our theory and demonstrate this learning scheme significantly improves performance.
****
On Explaining Neural Network Robustness with Activation Path
Ziping Jiang
Despite their verified performance, neural networks are prone to be misled by maliciously designed adversarial examples. This work investigates the robustness of neural networks from the activation pattern perspective. We find that despite the complex structure of the deep neural network, most of the neurons provide locally stable contributions to the output, while the minority, which we refer to as float neurons, can greatly affect the prediction. We decompose the computational graph of the neural network into the fixed path and float path and investigate their role in generating adversarial examples. Based on our analysis, we categorize the vulnerable examples into Lipschitz vulnerability and float neuron vulnerability. We show that the boost of robust accuracy from randomized smoothing is the result of correcting the latter. We then propose an SC-RFP (smoothed clas

sifier with repressed float path) to further reduce the instability of the float neurons and show that our result can provide a higher certified radius as well as accuracy.

****

Flareon: Stealthy Backdoor Injection via Poisoned Augmentation

Tianrui Qin,Xitong Gao,Xianghuan He,Yiren Zhao,Kejiang Ye,Cheng-zhong Xu

Open software supply chain attacks, once successful, can exact heavy costs in mission-critical applications. As open-source ecosystems for deep learning flourish and become increasingly universal, they present attackers previously unexplored avenues to code-inject malicious backdoors in deep neural network models. This paper proposes Flareon, a simple, stealthy, mostly-free, and yet effective backdoor injection payload that specifically targets the data augmentation pipeline with motion-based triggers. Flareon neither alters ground-truth labels, nor modifies the training loss objective, nor does it assume prior knowledge of the victim model architecture and training hyperparameters. By learning multiple triggers for targets simultaneously, it can even produce models that learn target-conditional (or ``any2any'') backdoors. Model trained under Flareon exhibits higher attack success rates for any target choices and better clean accuracies than competing attacks that not only seize greater capabilities, but also assume more restrictive attack targets. We also demonstrate the effectiveness of Flareon against recent defenses. Flareon is fully open-source and available online to the deep learning community.

****

Dimensionless instance segmentation by learning graph representations of point clouds

Robert Kiewisz,Tristan Bepler

Point clouds are an increasingly common spatial data modality, being produced by sensors used in robotics and self-driving cars, and as natural intermediate representations of objects in microscopy and other bioimaging domains (e.g., cell locations over time, or filaments, membranes, or organelle boundaries in cryo-electron micrographs or tomograms). However, semantic and instance segmentation of this data remains challenging due to the complex nature of objects in point clouds. Especially in bioimaging domains where objects are often large and can be intersecting or overlapping. Furthermore, methods for operating on point clouds should not be sensitive to the specific orientation or translation of the point cloud, which is often arbitrary. Here, we frame the point cloud instance segmentation problem as a graph learning problem in which we seek to learn a function that accepts the point cloud as input and outputs a probability distribution over neighbor graphs in which connected components of the graph correspond to individual object instances. We introduce the Dimensionless Instance Segmentation Transformer (DIST), a deep neural network for spatially invariant instance segmentation of point clouds to solve this point cloud-to-graph problem. DIST uses an SO(n) invariant transformer layer architecture to operate on point clouds of arbitrary dimension and outputs, for each pair of points, the probability that an edge exists between them in the instance graph. We then decode the most likely set of instances using a graph cut. We demonstrate the power of DIST for the segmentation of biomolecules in cryo-electron micrographs and tomograms, far surpassing existing methods for membrane and filament segmentation in empirical evaluation. DIST also applies to scene and object understanding, performing competitively on the ScanNetV2 3D instance segmentation challenge. We anticipate that DIST will underpin a new generation of methods for point cloud segmentation in bioimaging and that our general model and approach will provide useful insights for point cloud segmentation methods in other domains.

****

Structure by Architecture: Structured Representations without Regularization

Felix Leeb,Giulia Lanzillotta,Yashas Annadani,Michel Besserve,Stefan Bauer,Bernhard Schölkopf

We study the problem of self-supervised structured representation learning using autoencoders for downstream tasks such as generative modeling. Unlike most methods which rely on matching an arbitrary, relatively unstructured, prior distribu tion for sampling, we propose a sampling technique that relies solely on the independence of latent variables, thereby avoiding the trade-off between reconstruction quality and generative performance typically observed in VAEs. We design a novel autoencoder architecture capable of learning a structured representation without the need for aggressive regularization. Our structural decoders learn a hierarchy of latent variables, thereby ordering the information without any additional regularization or supervision. We demonstrate how these models learn a representation that improves results in a variety of downstream tasks including generation, disentanglement, and extrapolation using several challenging and natural image datasets.

****

Understanding Neural Coding on Latent Manifolds by Sharing Features and Dividing Ensembles

Martin Bjerke,Lukas Schott,Kristopher T Jensen,Claudia Battistin,David A. Klindt,Benjamin Adric Dunn

Systems neuroscience relies on two complementary views of neural data, characterized by single neuron tuning curves and analysis of population activity. These two perspectives combine elegantly in neural latent variable models that constrain the relationship between latent variables and neural activity, modeled by simple tuning curve functions. This has recently been demonstrated using Gaussian processes, with applications to realistic and topologically relevant latent manifolds. Those and previous models, however, missed crucial shared coding properties of neural populations. We propose $\textit{feature sharing}$ across neural tuning curves which significantly improves performance and helps optimization. We also propose a solution to the $\textit{ensemble detection}$ problem, where different groups of neurons, i.e., ensembles, can be modulated by different latent manifolds. Achieved through a soft clustering of neurons during training, this allows for the separation of mixed neural populations in an unsupervised manner. These innovations lead to more interpretable models of neural population activity that train well and perform better even on mixtures of complex latent manifolds. Finally, we apply our method on a recently published grid cell dataset, and recover distinct ensembles, infer toroidal latents and predict neural tuning curves in a single integrated modeling framework.

****

Learning Fast and Slow for Time Series Forecasting

Quang Pham,Chenghao Liu,Doyen Sahoo,Steven Hoi

Despite the recent success of deep learning for time series forecasting, these methods are not scalable for many real-world applications where data arrives sequentially. Training deep neural forecasters on the fly is notoriously challenging because of their limited ability to adapt to non-stationary environments and remember old knowledge. We argue that the fast adaptation capability of deep neural networks is critical and successful solutions require handling changes to both new and recurring patterns effectively. In this work, inspired by the Complementary Learning Systems (CLS) theory, we propose Fast and Slow learning Network (FSNet) as a novel framework to address the challenges of online forecasting. Particularly, FSNet improves the slowly-learned backbone by dynamically balancing fast adaptation to recent changes and retrieving similar old knowledge. FSNet achieves this mechanism via an interaction between two novel complementary components: (i) a per-layer adapter to support fast learning from individual layers, and (ii) an associative memory to support remembering, updating, and recalling repeating events. Extensive experiments on real and synthetic datasets validate FSNet's efficacy and robustness to both new and recurring patterns. Our code is publicly available at: \url{https://github.com/salesforce/fsnet/}.

****

Perturbation Defocusing for Adversarial Defense

HENG YANG,Ke Li

Recent research indicates adversarial attacks are likely to deceive neural systems, including large-scale, pre-trained language models. Given a natural sentence, an attacker replaces a subset of words to fool objective models. To defend against adversarial attacks, existing works aim to reconstruct the adversarial exam ples. However, these methods show limited defense performance on the adversarial examples whilst also damaging the clean performance on natural examples. To achieve better defense performance, our finding indicates that the reconstruction of adversarial examples is not necessary. More specifically, we inject non-toxic perturbations into adversarial examples, which can disable almost all malicious perturbations. In order to minimize performance sacrifice, we employ an adversarial example detector to distinguish and repair detected adversarial examples, which alleviates the mis-defense on natural examples. Our experimental results on three datasets, two objective models and a variety of adversarial attacks show that the proposed method successfully repairs up to ~ 97% correctly identified adversarial examples with ≤~ 2% performance sacrifice. We provide an anony-mus demonstration of adversarial detection and repair based on our work.

****

Accuracy Boosters: Epoch-Driven Mixed-Mantissa Block Floating-Point for DNN Training

Simla Burcu Harma,Canberk Sönmez,Babak Falsafi,Martin Jaggi,Yunho Oh

The unprecedented growth in DNN model complexity, size and the amount of training data have led to a commensurate increase in demand for computing and a search for minimal encoding. Recent research advocates Hybrid Block Floating-Point (HBFP) as a technique that minimizes silicon provisioning in accelerators by converting the majority of arithmetic operations in training to 8-bit fixed-point. In this paper, we perform a full-scale exploration of the HBFP design space including minimal mantissa encoding, varying block sizes, and mixed mantissa bit-width across layers and epochs. We propose Accuracy Boosters, an epoch-driven mixed-mantissa HBFP that uses 6-bit mantissa only in the last epoch and converts 99.7% of all arithmetic operations in training to 4-bit mantissas. Accuracy Boosters enable reducing silicon provisioning for an HBFP training accelerator by 16.98× as compared to FP32, while preserving or outperforming FP32 accuracy.

****

Compressing multidimensional weather and climate data into neural networks

Langwen Huang,Torsten Hoefler

Weather and climate simulations produce petabytes of high-resolution data that are later analyzed by researchers in order to understand climate change or severe weather. We propose a new method of compressing this multidimensional weather and climate data: a coordinate-based neural network is trained to overfit the data, and the resulting parameters are taken as a compact representation of the original grid-based data. While compression ratios range from 300x to more than 3,000x, our method outperforms the state-of-the-art compressor SZ3 in terms of weighted RMSE, MAE. It can faithfully preserve important large scale atmosphere structures and does not introduce significant artifacts.

When using the resulting neural network as a 790x compressed dataloader to train the WeatherBench forecasting model, its RMSE increases by less than 2%. The three orders of magnitude compression democratizes access to high-resolution climate data and enables numerous new research directions.

****

Guess the Instruction! Flipped Learning Makes Language Models Stronger Zero-Shot Learners

Seonghyeon Ye,Doyoung Kim,Joel Jang,Joongbo Shin,Minjoon Seo

Meta-training, which fine-tunes the language model (LM) on various downstream tasks by maximizing the likelihood of the target label given the task instruction and input instance, has improved the zero-shot task generalization performance. However, meta-trained LMs still struggle to generalize to challenging tasks containing novel labels unseen during meta-training. In this paper, we propose Flipped Learning, an alternative method of meta-training which trains the LM to generate the task instruction given the input instance and label. During inference, the LM trained with Flipped Learning, referred to as FLIPPED, selects the label option that is most likely to generate the task instruction. On 14 tasks of the BIG-bench benchmark, the 11B-sized FLIPPED outperforms zero-shot T0-11B (Sanh et al, 2021) and even a 16 times larger 3-shot GPT-3 (175B) (Brown et al, 2020) on average by 8.4% and 9.7% points, respectively. FLIPPED gives particularly large improvements on tasks with unseen labels, outperforming T0-11B by up to +20% average F1 score. This indicates that the strong task generalization of FLIPPED comes from improved generalization to novel labels. We release our code at github.com/seonghyeonye/Flipped-Learning.

****

Probabilistic Imputation for Time-series Classification with Missing Data

Hyunsu Kim,SeungHyun Kim,Eunggu Yun,Hwangrae Lee,Jaehun Lee,Juho Lee

Multivariate time series data available for real-world applications typically contain a significant amount of missing values. A dominant approach for the classification with such missing values is to heuristically impute the missing values with specific values (zero, mean, values of adjacent time-steps) or learnable parameters. However, these simple strategies do not take the data generative process into account, and more importantly, do not effectively capture the uncertainty in prediction due to the multiple possibilities for the missing values. In this paper, we propose a novel probabilistic framework for classification with multivariate time series data with missing values. Our model consists of two parts; a deep generative model for missing value imputation and a classifier. Extending the existing deep generative models to better capture structures of time-series data, our deep generative model part is trained to impute the missing values in multiple plausible ways, effectively modeling the uncertainty of the imputation. The classifier part takes the time series data along with the imputed missing values and classifies signals, and is trained to capture the predictive uncertainty due to the multiple possibilities of imputations. Importantly, we show that na\"ively combining the generative model and the classifier could result in trivial solutions where the generative model does not produce meaningful imputations. To resolve this, we present a novel regularization technique that can promote the model to produce useful imputation values that actually help classification. Through extensive experiments on real-world time series data with missing values, we demonstrate the effectiveness of our method.

****

Timing is Everything: Learning to Act Selectively with Costly Actions and Budgetary Constraints

David Henry Mguni,Aivar Sootla,Juliusz Krzysztof Ziomek,Oliver Slumbers,Zipeng Dai,Kun Shao,Jun Wang

Many real-world settings involve costs for performing actions; transaction costs in financial systems and fuel costs being common examples. In these settings, performing actions at each time step quickly accumulates costs leading to vastly suboptimal outcomes. Additionally, repeatedly acting produces wear and tear and ultimately, damage. Determining when to act is crucial for achieving successful outcomes and yet, the challenge of efficiently learning to behave optimally when actions incur minimally bounded costs remains unresolved. In this paper, we introduce a reinforcement learning (RL) framework named Learnable Impulse Control Reinforcement Algorithm (LICRA), for learning to optimally select both when to act and which actions to take when actions incur costs. At the core of LICRA is a nested structure that combines RL and a form of policy known as impulse control which learns to maximise objectives when actions incur costs. We prove that LICRA, which seamlessly adopts any RL method, converges to policies that optimally select when to perform actions and their optimal magnitudes. We then augment LICRA to handle problems in which the agent can perform at most $k < \infty$ actions and more generally, faces a budget constraint. We show LICRA learns the optimal value function and ensures budget constraints are satisfied almost surely.
We demonstrate empirically LICRA’s superior performance against benchmark RL methods in OpenAI gym’s Lunar Lander and in Highway environments and a variant of the Merton portfolio problem within finance.

****

Multi-Head State Space Model for Sequence Modeling

Yassir Fathullah,Chunyang Wu,Yuan Shangguan,Junteng Jia,Wenhan Xiong,Jay Mahadeokar,Chunxi Liu,Yangyang Shi,Ozlem Kalinli,Mike Seltzer,Mark Gales

Recently, state space models (SSMs) have shown promising results on sequence modeling tasks. However, a potential challenge of existing works is that SSMs are usually introduced or initialized in a homogeneous way, encouraging the model to only capture similar temporal dynamics on different features. In this paper, we propose a multi-head state space model (MSSM), in which parallel heads are introduced to learn different temporal dynamics on sequence data. Furthermore, we propose a novel variant of the Transformer, referred to as the Stateformer, which combines MSSMs with attention. Experiments on large-scale automatic speech recognition (ASR) and language modeling tasks show the MSSM outperforming a range of attention-based baselines. The Stateformer further improves performance, achieving the state-of-the-art performance on the LibriSpeech ASR task.
****
A Weight Variation-Aware Training Method for Hardware Neuromorphic Chips
Min-Hye Oh
Hardware neuromorphic chips that mimic the biological nervous systems have recently attracted significant attention due to their ultra-low power and parallel computation. However, the inherent variability of nano-scale synaptic devices causes a weight perturbation and performance drop of neural networks. This paper proposes a training method to find weight with robustness to intrinsic device variability. A stochastic weight characteristic incurred by device inherent variability is considered during training. We investigate the impact of weight variation on both Spiking Neural Network (SNN) and standard Artificial Neural Network (ANN) with different architectures including fully connected, convolutional neural network (CNN), VGG, and ResNet on MNIST, CIFAR-10, and CIFAR-100. Experimental results show that a weight variation-aware training method (WVAT) can dramatically minimize the performance drop on weight variability by exploring a flat loss landscape. When there are weight perturbations, WVAT yields 85.21% accuracy of VGG-5 on CIFAR-10, reducing accuracy degradation by more than 1/10 compared with SGD. Finally, WVAT is easy to implement on various architectures with little computational overhead.
****
Semantic Prior for Weakly Supervised Class-Incremental Segmentation
Subhankar Roy,Riccardo Volpi,Gabriela Csurka,Diane Larlus
Class-incremental semantic image segmentation assumes multiple model updates, each enriching the model to segment new categories. This is typically carried out by providing pixel-level manual annotations for all new objects, limiting the adoption of such methods. Approaches which solely require image-level labels offer an attractive alternative, yet, such annotations lack crucial information about the location and boundary of new objects. In this paper we argue that, since classes represent not just indices but semantic entities, the conceptual relationships between them can provide valuable information that should be leveraged. We propose a weakly supervised approach that leverages such semantic relations in order to transfer some cues from the previously learned classes into the new ones, complementing the supervisory signal from image-level labels. We validate our approach on a number of continual learning tasks, and show how even a simple pairwise interaction between classes can significantly improve the segmentation mask quality of both old and new classes. We show these conclusions still hold for longer and, hence, more realistic sequences of tasks and for a challenging few-shot scenario.
****
A Mutual Information Duality Algorithm for Multi-Agent Specialization
Stefan Juang,Qiyang Cao,Yuan Zhou,Ruochen Liu,Nevin Zhang,Elvis S. Liu
The social behavior change in a population has long been studied as an essential component of multi-agent learning. The learning of behavioral change not only involves reinforcement learning (RL), but also be measured against the general population with mutual information (MI). The combination of RL and MI led us to derive MI optimizations from policy gradient. With MI as multi-agent's optimization objective, we discover that the dual properties of MI can result in distinctly different population behaviors. From MI maximization that maximizes the stability of a population to MI minimization that enables specialization among the agen

ts, the dual of MI creates a significant change in a population's behavioral properties. In this paper, we propose a minimax formulation of MI (M\&M) that enables agents specialization with stable regularization. Empirically we evaluated M\&M against the prior SOTA MARL framework, and analyze the social behavior change in performance, diversity, and the stability of their social graphs.

****

DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training
Wei Li,Linchao Zhu,Longyin Wen,Yi Yang

Large-scale pre-trained multi-modal models (e.g., CLIP) demonstrate strong zero-shot transfer capability in many discriminative tasks, e.g., image classification. Their adaptation to zero-shot image-conditioned text generation tasks has drawn increasing interest. Prior arts approach to zero-shot captioning by either utilizing the existing large language models (e.g., GPT-2) or pre-training the encoder-decoder network in an end-to-end manner. However, the large language models may not generate sensible descriptions due to the task discrepancy between captioning and language modeling, while the end-to-end pre-training requires paired data and extensive computational resources. In this work, we propose a simple framework, named DeCap, for zero-shot captioning. We introduce a lightweight visual-aware language decoder. This decoder is both data-efficient and computation-efficient: 1) it only requires the \textit{text} data for training, easing the burden on the collection of paired data. 2) it does not require end-to-end training. When trained with text-only data, the decoder takes the text embedding extracted from the off-the-shelf CLIP encoder as a prefix embedding. The challenge is that the decoder is trained on the text corpus but at the inference stage, it needs to generate captions based on visual inputs. Though the CLIP text embedding and the visual embedding are correlated, the \textit{modality gap} issue is widely observed in multi-modal contrastive models that prevents us from directly taking the visual embedding as the prefix embedding. We propose a training-free mechanism to reduce the modality gap. We project the visual embedding into the CLIP text embedding space, while the projected embedding retains the information of the visual input. Taking the projected embedding as the prefix embedding, the decoder generates high-quality descriptions that match the visual input. The experiments show that DeCap outperforms other zero-shot captioning methods and unpaired captioning methods by a large margin on the typical image captioning benchmarks, i.e., MSCOCO and NoCaps. We apply DeCap to video captioning and achieve state-of-the-art zero-shot performance on MSR-VTT and ActivityNet-Captions. The code is available at https://github.com/dhg-wei/DeCap.

****

Biological Factor Regulatory Neural Network
Xinnan Dai,Caihua Shan,Jie Zheng,Xiaoxiao Li,Dongsheng Li

Genes are fundamental for analyzing biological systems and many recent works proposed to utilize gene expression for various biological tasks by deep learning models. Despite their promising performance, it is hard for deep neural networks to provide biological insights for humans due to their black-box nature. Recently, some works integrated biological knowledge with neural networks to improve the transparency and performance of their models. However, these methods can only incorporate partial biological knowledge, leading to suboptimal performance. In this paper, we propose the Biological Factor Regulatory Neural Network (BFReg-NN), a generic framework to model relations among biological factors in cell systems. BFReg-NN starts from gene expression data and is capable of merging most existing biological knowledge into the model, including the regulatory relations among genes or proteins (e.g., gene regulatory networks (GRN), protein-protein interaction networks (PPI)) and the hierarchical relations among genes, proteins and pathways (e.g., several genes/proteins are contained in a pathway). Moreover, BFReg-NN also has the ability to provide new biologically meaningful insights because of its white-box characteristics. Experimental results on different gene expression-based tasks verify the superiority of BFReg-NN compared with baselines. Our case studies also show that the key insights found by BFReg-NN are consistent with the biological literature.

****

Preserving Semantics in Textual Adversarial Attacks
David Herel,Hugo Cisneros,Tomas Mikolov
Adversarial attacks in NLP challenge the way we look at language models. The goal of this kind of adversarial attack is to modify the input text to fool a classifier while maintaining the original meaning of the text. Although most existing adversarial attacks claim to fulfill the constraint of semantics preservation, careful scrutiny shows otherwise. We show that the problem lies in the text encoders used to determine the similarity of adversarial examples, specifically in the way they are trained. Unsupervised training methods make these encoders more susceptible to problems with antonym recognition. To overcome this, we introduce a simple, fully supervised sentence embedding technique called Semantics-Preserving-Encoder (SPE). The results show that our solution minimizes the variation in the meaning of the adversarial examples generated. It also significantly improves the overall quality of adversarial examples, as confirmed by human evaluators. Furthermore, it can be used as a component in any existing attack to speed up its execution while maintaining similar attack success.
****
Unbiased Decisions Reduce Regret: Adversarial Optimism for the Bank Loan Problem
Elena Gal,Shaun Singh,Aldo Pacchiano,Benjamin Walker,Terry Lyons,Jakob Nicolaus Foerster
In many real world settings binary classification decisions are made based on limited data in near real-time, e.g. when assessing a loan application. We focus on a class of these problems that share a common feature that the true label is only observed when a data point is assigned a positive label by a learner, e.g. we only learn of an outcome of \emph{accepted} loan applications. In this setting, sometimes referred to as the Bank Loan Problem (BLP) in the literature, the labelled training set suffers from accumulating bias since it is created by learners past decisions.
Prior work mitigates the consequences of this bias by injecting optimism into the model to allow the learner to correct self-reinforcing false rejections. This reduces long term regret but comes at the cost of a higher false acceptance rate.
We introduce \emph{adversarial optimism} (AdOpt) to directly address the bias in the training set using \emph{adversarial domain adaptation}. The goal of AdOpt is to learn an unbiased but informative representation of past data, by reducing the distributional shift between the set of \textit{accepted} data points and all data points seen thus far. AdOpt integrates classification made using this debiased representation of the data with the recently proposed \emph{pseudo-label optimism}(PLOT) method to increase the rate of correct decisions at every timestep.
AdOpt significantly exceeds state-of-the-art performance on a set of challenging BLP benchmark problems.
****
That Label's got Style: Handling Label Style Bias for Uncertain Image Segmentation
Kilian Zepf,Eike Petersen,Jes Frellsen,Aasa Feragen
Segmentation uncertainty models predict a distribution over plausible segmentations for a given input, which they learn from the annotator variation in the training set. However, in practice these annotations can differ systematically in the way they are generated, for example through the use of different labeling tools. This results in datasets that contain both data variability and differing label styles. In this paper, we demonstrate that applying state-of-the-art segmentation uncertainty models on such datasets can lead to model bias caused by the different label styles. We present an updated modelling objective conditioning on labeling style for aleatoric uncertainty estimation, and modify two state-of-the-art-architectures for segmentation uncertainty accordingly. We show with extensive experiments that this method reduces label style bias, while improving segmentation performance, increasing the applicability of segmentation uncertainty models in the wild. We curate two datasets, with annotations in different label styles, which we will make publicly available along with our code upon publication

.
****

Holistic Adversarially Robust Pruning
Qi Zhao,Christian Wressnegger
Neural networks can be drastically shrunk in size by removing redundant parameters. While crucial for the deployment on resource-constraint hardware, oftentimes, compression comes with a severe drop in accuracy and lack of adversarial robustness. Despite recent advances, counteracting both aspects has only succeeded for moderate compression rates so far. We propose a novel method, HARP, that copes with aggressive pruning significantly better than prior work. For this, we consider the network holistically. We learn a global compression strategy that optimizes how many parameters (compression rate) and which parameters (scoring connections) to prune specific to each layer individually. Our method fine-tunes an existing model with dynamic regularization, that follows a step-wise incremental function balancing the different objectives. It starts by favoring robustness before shifting focus on reaching the target compression rate and only then handles the objectives equally. The learned compression strategies allow us to maintain the pre-trained model’s natural accuracy and its adversarial robustness for a reduction by 99% of the network’s original size. Moreover, we observe a crucial influence of non-uniform compression across layers. The implementation of HARP is publicly available at https://intellisec.de/research/harp.

****

PASHA: Efficient HPO and NAS with Progressive Resource Allocation
Ondrej Bohdal,Lukas Balles,Martin Wistuba,Beyza Ermis,Cedric Archambeau,Giovanni Zappella
Hyperparameter optimization (HPO) and neural architecture search (NAS) are methods of choice to obtain the best-in-class machine learning models, but in practice they can be costly to run. When models are trained on large datasets, tuning them with HPO or NAS rapidly becomes prohibitively expensive for practitioners, even when efficient multi-fidelity methods are employed. We propose an approach to tackle the challenge of tuning machine learning models trained on large datasets with limited computational resources. Our approach, named PASHA, extends ASHA and is able to dynamically allocate maximum resources for the tuning procedure depending on the need. The experimental comparison shows that PASHA identifies well-performing hyperparameter configurations and architectures while consuming significantly fewer computational resources than ASHA.

****

Thinking fourth dimensionally: Treating Time as a Random Variable in EBMs
Omer Yair,Tomer Michaeli
Recent years have seen significant progress in techniques for learning high-dimensional distributions. Many modern methods, from diffusion models to Energy-Based-Models (EBMs), adopt a coarse-to-fine approach. This is often done by introducing a series of auxiliary distributions that gradually change from the data distribution to some simple distribution (e.g. white Gaussian noise). Methods in this category separately learn each auxiliary distribution (or transition between pairs of consecutive distributions) and then use the learned models sequentially to generate samples. In this paper, we offer a simple way to generalize this idea by treating the ``time'' index of the series as a random variable and framing the problem as that of learning a single joint distribution of "time" and samples. We show that this joint distribution can be learned using any existing EBM method and that it allows achieving improved results. As an example, we demonstrate this approach using contrastive divergence (CD) in its most basic form. On CIFAR-10 and CelebA (32×32), this method outperforms previous CD-based methods in terms of inception and FID scores.

****

Diversity of Generated Unlabeled Data Matters for Few-shot Hypothesis Adaptation
Ruijiang Dong,Feng Liu,Haoang Chi,Tongliang Liu,Mingming Gong,Gang Niu,Masashi Sugiyama,Bo Han
Generating unlabeled data has been recently shown to help address the few-shot hypothesis adaptation (FHA) problem, where we aim to train a classifier for the t

arget domain with a few labeled target-domain data and a well-trained source-domain classifier (i.e., a source hypothesis), for the additional information of the highly-compatible unlabeled data. However, the generated data of the existing methods are extremely similar or even the same. The strong dependency among the generated data will lead the learning to fail. In this paper, we propose a diversity-enhancing generative network (DEG-Net) for the FHA problem, which can generate diverse unlabeled data with the help of a kernel independence measure: the Hilbert-Schmidt independence criterion (HSIC). Specifically, DEG-Net will generate data via minimizing the HSIC value (i.e., maximizing the independence) among the semantic features of the generated data. By DEG-Net, the generated unlabeled data are more diverse and more effective for addressing the FHA problem. Experimental results show that the DEG-Net outperforms existing FHA baselines and further verifies that generating diverse data plays an important role in addressing the FHA problem.

****

StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random

Haoxuan Li,Chunyuan Zheng,Peng Wu

In recommender systems, users always choose the favorite items to rate, which leads to data missing not at random and poses a great challenge for unbiased evaluation and learning of prediction models. Currently, the doubly robust (DR) methods have been widely studied and demonstrate superior performance. However, in this paper, we show that DR methods are unstable and have unbounded bias, variance, and generalization bounds to extremely small propensities. Moreover, the fact that DR relies more on extrapolation will lead to suboptimal performance. To address the above limitations while retaining double robustness, we propose a stabilized doubly robust (StableDR) learning approach with a weaker reliance on extrapolation. Theoretical analysis shows that StableDR has bounded bias, variance, and generalization error bound simultaneously under inaccurate imputed errors and arbitrarily small propensities. In addition, we propose a novel learning approach for StableDR that updates the imputation, propensity, and prediction models cyclically, achieving more stable and accurate predictions. Extensive experiments show that our approaches significantly outperform the existing methods.

****

Variational Causal Dynamics: Discovering Modular World Models from Interventions

Anson Lei,Bernhard Schölkopf,Ingmar Posner

Latent world models allow agents to reason about complex environments with high-dimensional observations. However, adapting to new environments and effectively leveraging previous knowledge remain significant challenges. We present variational causal dynamics (VCD), a structured world model that exploits the invariance of causal mechanisms across environments to achieve fast and modular adaptation. By causally factorising a transition model, VCD is able to identify reusable components across different environments. This is achieved by combining causal discovery and variational inference to learn a latent representation and transition model jointly in an unsupervised manner. Specifically, we optimise the evidence lower bound jointly over a representation model and a transition model structured as a causal graphical model. In evaluations on simulated environments with state and image observations, we show that VCD is able to successfully identify causal variables, and to discover consistent causal structures across different environments. Moreover, given a small number of observations in a previously unseen, intervened environment, VCD is able to identify the sparse changes in the dynamics and to adapt efficiently. In doing so, VCD significantly extends the capabilities of the current state-of-the-art in latent world models while also comparing favourably in terms of prediction accuracy.

****

Query The Agent: Improving Sample Efficiency Through Epistemic Uncertainty Estimation

Julian Alverio,Boris Katz,Andrei Barbu

Curricula for goal-conditioned reinforcement learning agents typically rely on poor estimates of the agent's epistemic uncertainty or fail to consider the agent s' epistemic uncertainty altogether, resulting in poor sample efficiency. We propose a novel algorithm, Query The Agent (QTA), which significantly improves sample efficiency by estimating the agent's epistemic uncertainty throughout the state space and setting goals in highly uncertain areas. Encouraging the agent to collect data in highly uncertain states allows the agent to improve its estimation of the value function rapidly. QTA utilizes a novel technique for estimating epistemic uncertainty, Predictive Uncertainty Networks (PUN), to allow QTA to assess the agent's uncertainty in all previously observed states. We demonstrate that QTA offers decisive sample efficiency improvements over preexisting methods.

****

Differentiable Logic Programming for Probabilistic Reasoning

Tuo Xu,Lei Zou

This paper studies inductive logic programming for probabilistic reasoning. The key problems, i.e. learning rule structures and learning rule weights, have been extensively studied with traditional discrete searching methods as well as recent neural-based approaches. In this paper, we present a new approach called Differentiable Logic Programming (DLP), which provides a flexible framework for learning first-order logical rules for reasoning. We propose a continuous version of optimization problem for learning high-quality rules as a proxy and generalize rule learning and forward chaining algorithms in a differentiable manner, which enables us to efficiently learn rule structures and weights via gradient-based methods. Theoretical analysis and empirical results show effectiveness of our approach.

****

Rewiring with Positional Encodings for GNNs

Rickard Brüel Gabrielsson,Mikhail Yurochkin,Justin Solomon

Several recent works use positional encodings to extend the receptive fields of graph neural network (GNN) layers equipped with attention mechanisms. These techniques, however, extend receptive fields to the complete graph, at substantial computational cost and risking a change in the inductive biases of conventional GNNs, or require complex architecture adjustments. As a conservative alternative, we use positional encodings to expand receptive fields to r-hop neighborhoods. More specifically, our method augments the input graph with additional nodes/edges and uses positional encodings as node and/or edge features. We thus modify graphs before inputting them to a downstream GNN model, instead of modifying the model itself. This makes our method model-agnostic, i.e. compatible with any existing GNN architectures. We also provide examples of positional encodings that are lossless with a one-to-one map between the original and the modified graphs. We demonstrate that extending receptive fields via positional encodings and a virtual fully-connected node significantly improves GNN performance and alleviates over-squashing using small r. We obtain improvements on a variety of models and datasets, and reach state-of-the-art performance using traditional GNNs or graph Transformers.

****

Feed-Forward Latent Domain Adaptation

Ondrej Bohdal,Da Li,Shell Xu Hu,Timothy Hospedales

We study the highly practical but comparatively under-studied problem of latent-domain adaptation, where a source model should be adapted to a target dataset that contains a mixture of unlabelled domain-relevant and domain-irrelevant examples. Furthermore, motivated by the requirements for data privacy and the need for embedded and resource-constrained devices of all kinds to adapt to local data distributions, we focus on the setting of feed-forward source-free domain adaptation, where adaptation should not require access to the source dataset, and also be back propagation-free. Our solution is to meta-learn a network capable of embedding the mixed-relevance target dataset and dynamically adapting inference for target examples using cross-attention. The resulting framework leads to consistent improvement on strong ERM baselines. We also show that our framework sometimes even improves on the upper bound of domain-supervised adaptation, where only domain-relevant instances are provided for adaptation. This suggests that human annotated domain labels may not always be optimal, and raises the possibility o f doing better through automated instance selection.

****

Sampling-based inference for large linear models, with application to linearised Laplace

Javier Antoran,Shreyas Padhy,Riccardo Barbano,Eric Nalisnick,David Janz,José Miguel Hernández-Lobato

Large-scale linear models are ubiquitous throughout machine learning, with contemporary application as surrogate models for neural network uncertainty quantification; that is, the linearised Laplace method. Alas, the computational cost associated with Bayesian linear models constrains this method's application to small networks, small output spaces and small datasets. We address this limitation by introducing a scalable sample-based Bayesian inference method for conjugate Gaussian multi-output linear models, together with a matching method for hyperparameter (regularisation) selection. Furthermore, we use a classic feature normalisation method (the g-prior) to resolve a previously highlighted pathology of the linearised Laplace method. Together, these contributions allow us to perform linearised neural network inference with ResNet-18 on CIFAR100 (11M parameters, 100 output dimensions × 50k datapoints) and with a U-Net on a high-resolution tomographic reconstruction task (2M parameters, 251k output dimensions).

****

Defending against Adversarial Audio via Diffusion Model

Shutong Wu,Jiongxiao Wang,Wei Ping,Weili Nie,Chaowei Xiao

Deep learning models have been widely used in commercial acoustic systems in recent years. However, adversarial audio examples can cause abnormal behaviors for those acoustic systems, while being hard for humans to perceive. Various methods, such as transformation-based defenses and adversarial training, have been proposed to protect acoustic systems from adversarial attacks, but they are less effective against adaptive attacks. Furthermore, directly applying the methods from the image domain can lead to suboptimal results because of the unique properties of audio data. In this paper, we propose an adversarial purification-based defense pipeline, AudioPure, for acoustic systems via off-the-shelf diffusion models. Taking advantage of the strong generation ability of diffusion models, AudioPure first adds a small amount of noise to the adversarial audio and then runs the reverse sampling step to purify the noisy audio and recover clean audio. AudioPure is a plug-and-play method that can be directly applied to any pretrained classifier without any fine-tuning or re-training. We conduct extensive experiments on the speech command recognition task to evaluate the robustness of AudioPure. Our method is effective against diverse adversarial attacks (e.g. L2 or L∞-norm). It outperforms the existing methods under both strong adaptive white-box and black-box attacks bounded by L2 or L∞-norm (up to +20% in robust accuracy). Besides, we also evaluate the certified robustness for perturbations bounded by L2-norm via randomized smoothing. Our pipeline achieves a higher certified accuracy than baselines.

****

Text-Guided Diffusion Image Style Transfer with Contrastive Loss Fine-tuning

Serin Yang,Hyunmin Hwang,Jong Chul Ye

Recently, diffusion models have demonstrated superior performance in text-guided image style transfer. However, there exists fundamental trade-off between transforming styles and maintaining content in diffusion models. Although a simple remedy would be using deterministic sampling scheme such as denoising diffusion implicit model (DDIM) that guarantees the perfect reconstruction, it requires the computationally expensive fining-tuning of the diffusion models. To address this, here we present a text-guided sampling scheme using a patch-wise contrastive loss fine-tuning. By exploiting the contrastive loss between the samples and the original images, our diffusion model can generate an image with the same semantic content as the source image. Experimental results demonstrate that our approach outperforms the existing methods while maintaining content and requiring no additional training on the diffusion model.

****

FedProp: Cross-client Label Propagation for Federated Semi-supervised Learning

Jonathan Scott,Michelle Yeo,Christoph H Lampert

Federated learning (FL) allows multiple clients to jointly train a machine learning model in such a way that no client has to share their data with any other participating party. In the supervised setting, where all client data is fully labeled, FL has been widely adopted for learning tasks that require data privacy. However, it is an ongoing research question how to best perform federated learning in a semi-supervised setting, where the clients possess data that is only partially labeled or even completely unlabeled. In this work, we propose a new method, FedProp, that follows a manifold-based approach to semi-supervised learning (SSL). It estimates the data manifold jointly from the data of multiple clients and computes pseudo-labels using cross-client label propagation. To avoid that clients have to share their data with anyone, FedProp employs two cryptographically secure yet highly efficient protocols: multi-party Hamming distance computation and secure aggregation. Experiments on three standard benchmarks show that FedProp achieves higher classification accuracy than previous federated SSL methods. Furthermore, as a pseudo-label-based technique, FedProp is complementary to other federated SSL approaches, in particular consistency-based ones. We demonstrate experimentally that further accuracy gains are possible by combining both.

****

Gated Inference Network: Inferencing and Learning State-Space Models

Hamidreza Hashempoor,Wan Choi

State-space models (SSMs) perform predictions by learning the underlying dynamics of observed sequence. We propose a new SSM in both high and low dimensional observation space, which utilizes Bayesian filtering-smoothing to model system’s dynamics more accurately than RNN-based SSMs and can be learned in an end-to-end manner. The designed architecture, which we call the Gated Inference Network (GIN), is able to integrate the uncertainty estimates and learn the complicated dynamics of the system that enables us to perform estimation and imputation tasks in both data presence and absence. The proposed model uses the GRU cells into its structure to complete the data flow, while avoids expensive computations and potentially unstable matrix inversions. The GIN is able to deal with any time-series data and gives us a strong robustness to handle the observational noise. In the numerical experiments, we show that the GIN reduces the
uncertainty of estimates and outperforms its counterparts , LSTMs, GRUs and variational approaches.

****

Theoretical Characterization of the Generalization Performance of Overfitted Meta-Learning

Peizhong Ju,Yingbin Liang,Ness Shroff

Meta-learning has arisen as a successful method for improving training performance by training over many similar tasks, especially with deep neural networks (DNNs). However, the theoretical understanding of when and why overparameterized models such as DNNs can generalize well in meta-learning is still limited. As an initial step towards addressing this challenge, this paper studies the generalization performance of overfitted meta-learning under a linear regression model with Gaussian features. In contrast to a few recent studies along the same line, our framework allows the number of model parameters to be arbitrarily larger than the number of features in the ground truth signal, and hence naturally captures the overparameterized regime in practical deep meta-learning. We show that the overfitted min ℓ_2-norm solution of model-agnostic meta-learning (MAML) can be beneficial, which is similar to the recent remarkable findings on "benign overfitting" and "double descent" phenomenon in the classical (single-task) linear regression. However, due to the uniqueness of meta-learning such as task-specific gradient descent inner training and the diversity/fluctuation of the ground-truth signals among training tasks, we find new and interesting properties that do not exist in single-task linear regression. We first provide a high-probability upper bound (under reasonable tightness) on the generalization error, where certain terms decrease when the number of features increases. Our analysis suggests that benign overfitting is more significant and easier to observe when the noise and the diversity/fluctuation of the ground truth of each training task are la rge. Under this circumstance, we show that the overfitted min ℓ_2-norm solution can achieve an even lower generalization error than the underparameterized solution.

****

Cold Posteriors through PAC-Bayes

Konstantinos Pitas,Julyan Arbel

We investigate the cold posterior effect through the lens of PAC-Bayes generalization bounds. We argue that in the non-asymptotic setting, when the number of training samples is (relatively) small, discussions of the cold posterior effect should take into account that approximate Bayesian inference does not readily provide guarantees of performance on out-of-sample data. Instead, out-of-sample error is better described through a generalization bound. In this context, we explore the connections of the ELBO objective from variational inference and the PAC-Bayes objectives. We note that, while the ELBO and PAC-Bayes objectives are similar, the latter objectives naturally contain a temperature parameter λ which is not restricted to be $\lambda=1$. For both simplified regression and realistic classification tasks, in the case of Laplace approximations to the posterior, we show how this PAC-Bayesian interpretation of the temperature parameter captures important aspects of the cold posterior effect.

****

Training language models to summarize narratives improves brain alignment

Khai Loong Aw,Mariya Toneva

Building systems that achieve a deeper understanding of language is one of the central goals of natural language processing (NLP). Towards this goal, recent works have begun to train language models on narrative datasets which require extracting the most critical information by integrating across long contexts. However, it is still an open question whether these models are learning a deeper understanding of the text, or if the models are simply learning a heuristic to complete the task. This work investigates this further by turning to the one language processing system that truly understands complex language: the human brain. We show that training language models for deeper narrative understanding results in richer representations that have improved alignment to human brain activity. We further find that the improvements in brain alignment are larger for character names than for other discourse features, which indicates that these models are learning important narrative elements. Taken together, these results suggest that this type of training can indeed lead to deeper language understanding. These findings have consequences both for cognitive neuroscience by revealing some of the significant factors behind brain-NLP alignment, and for NLP by highlighting that understanding of long-range context can be improved beyond language modeling.

****

VQ-TR: Vector Quantized Attention for Time Series Forecasting

Kashif Rasul,Umang Gupta,Hena Ghonia,Yuriy Nevmyvaka

Modern time series datasets can easily contain hundreds or thousands of temporal time points, however, Transformer based models scale poorly to the size of the sequence length constraining their context size in the seq-to-seq setting. In this work, we introduce VQ-TR which maps large sequences to a discrete set of latents representations as part of the Attention module. This allows us to attend over larger context windows with linear complexity with respect to the sequence length. We compare this method with other competitive deep learning and classical univariate probabilistic models and highlight its performance using both probabilistic and point forecasting metrics on a variety of open datasets from different domains.

****

Local KL Convergence Rate for Stein Variational Gradient Descent with Reweighted Kernel

Xunpeng Huang,Hanze Dong,Cong Fang

We study convergence properties of Stein Variational Gradient Descent (SVGD) algorithm for sampling from a non-normalized probabilistic distribution $p_*({x})\propto\exp(-f_*({x}))$. Compared with Kernelized Stein Discrepancy (KSD) convergence analyzed in previous literature, KL convergence as a more convincing criter ion can better explain the effectiveness of SVGD in real-world applications. In the population limit, SVGD performs smoothed gradient descent with kernel integral operator. Notably, SVGD with smoothing kernels suffers from gradient vanishing in low-density areas, which makes the error term between smoothed gradient and Wasserstein gradient not controllable. In this context, we introduce a reweighted kernel to amplify the smoothed gradient in low-density areas, which leads to a bounded error term to Wasserstein gradient. When the $p_*({x})$ satisfies log-Sobolev inequality, we develop the convergence rate for SVGD in KL divergence with the reweighted kernel. Our analysis points out the defects of conventional smoothing kernels in SVGD and provides the convergence rate for SVGD in KL divergence.

****

A Decomposition Based Dual Projection Model for Multivariate Time Series Forecasting and Anomaly Detection

Peng Zhang,Qin Xie,Jaesik Choi

Efficient anomaly detection and diagnosis in multivariate time series data is of great importance for various application areas. Forecasting of long-sequence time series is an important problem to prepare for future changes. An accurate prediction can help to detect anomaly events beforehand and make better decisions. It seems that one has to use more complex structures for deep learning models to get better performance, e.g., the recent surge of Transformer variants for time series modeling. However, such complex architectures require a large amount of training data and extensive computing resources. In addition, many of the considerations behind such architectures do not hold for time series applications. The objective of this study is to re-consider the effectiveness of deep learning architectures for efficient and accurate time series forecasting and anomaly detection. A model with direct projections is proposed, and it outperforms existing Transformer based models in most cases by a significant margin. The new decomposition based dual projection (DBDP) model consists of an anchored global profile and a varied number of decomposed seasonal local profiles of the time series for better forecasting performance. In addition to forecasting, a non-contrastive self-supervised learning approach, we propose to include a contrastive learning module in the DBDPC model for better forecasting performance and robustness. Finally, we apply the DBDP and DBDPC models to forecasting based time series anomaly detection and achieve superior performance over the latest SoTA models. These results demonstrate the effectiveness of the several key considerations behind the DBDP and DBDPC models, which also encourages the development of new architectures for time series applications.

****

FedHPO-Bench: A Benchmark Suite for Federated Hyperparameter Optimization

Zhen WANG,Weirui Kuang,Ce Zhang,Bolin Ding,Yaliang Li

Hyperparameter optimization (HPO) is crucial for machine learning algorithms to achieve satisfactory performance. Its research progress has been boosted by existing HPO benchmarks. Nonetheless, existing efforts in benchmarking all focus on HPO for traditional centralized learning while ignoring federated learning (FL), a promising paradigm for collaboratively learning models from dispersed data. In this paper, we first identify some uniqueness of HPO for FL algorithms from various aspects. Due to this uniqueness, existing HPO benchmarks no longer satisfy the need to compare HPO methods in the FL setting. To facilitate the research of HPO in the FL setting, we propose and implement a benchmark suite FedHPO-Bench that incorporates comprehensive FedHPO problems, enables flexible customization of the function evaluations, and eases continuing extensions. We also conduct extensive experiments based on FedHPO-Bench to provide the community with more insights into FedHPO. We open-sourced FedHPO-Bench at https://github.com/FedHPO-Bench/FedHPO-Bench-ICLR23.

****

CCT: Cross-consistency training for Clone Detection and Code Search Tasks

Nikita Sorokin,Dmitry Abulkhanov,Valentin Malykh

Clone detection is a well known task, which could be formulated on any programming language. Although to the best of our knowledge there is no cross-lingual clo ne detection task formulated. In this work we formulate such a task alongside with a specific training procedure CCT for a deep leaning language model. This procedure allows CCT-trained model to outperform the existing approaches on POJ-104 benchmark with result of 95.67\% MAP and on newly created cross-lingual clone detection benchmark XCD. Moreover, CCT model shows new state of the art results in code search task AdvTest 47.15\% MRR.

****

Robust Explanation Constraints for Neural Networks

Matthew Robert Wicker,Juyeon Heo,Luca Costabello,Adrian Weller

Post-hoc explanation methods are used with the intent of providing insights about neural networks and are sometimes said to help engender trust in their outputs. However, popular explanations methods have been found to be fragile to minor perturbations of input features or model parameters. Relying on constraint relaxation techniques from non-convex optimization, we develop a method that upper-bounds the largest change an adversary can make to a gradient-based explanation via bounded manipulation of either the input features or model parameters. By propagating a compact input or parameter set as symbolic intervals through the forwards and backwards computations of the neural network we can formally certify the robustness of gradient-based explanations. Our bounds are differentiable, hence we can incorporate provable explanation robustness into neural network training. Empirically, our method surpasses the robustness provided by previous heuristic approaches. We find that our training method is the only method able to learn neural networks with certificates of explanation robustness across all six datasets tested.

****

Cyclophobic Reinforcement Learning

Stefan Sylvius Wagner,Peter Arndt,Jan Robine,Stefan Harmeling

In environments with sparse rewards finding a good inductive bias for exploration is crucial to the agent’s success. However, there are two competing goals: novelty search and systematic exploration. While existing approaches such as curiousity- driven exploration find novelty, they sometimes do not systematically explore the whole state space, akin to depth-first-search vs breadth-first-search. In this paper, we propose a new intrinsic reward that is cyclophobic, i.e. it does not reward novelty, but punishes redundancy by avoiding cycles. Augmenting the cyclophobic intrinsic reward with a sequence of hierarchical representations based on the agent’s cropped observations we are able to achieve excellent results in the MiniGrid and MiniHack environments. Both are particularly hard, as they require complex interactions with different objects in order to be solved. Detailed comparisons with previous approaches and thorough ablation studies show that our newly proposed cyclophobic reinforcement learning is vastly more efficient than other state of the art methods.

****

Emergent collective intelligence from massive-agent cooperation and competition

Hanmo Chen,Stone Tao,Jiaxin Chen,Weihan Shen,Xihui Li,Sikai Cheng,Xiaolong Zhu,Xiu Li

Inspired by organisms evolving through cooperation and competition between different populations on Earth, we study the emergence of artificial collective intelligence through massive-agent reinforcement learning. To this end, We propose a new massive-agent reinforcement learning environment, Lux, where dynamic and massive agents in two teams scramble for limited resources and fight off the darkness. In Lux, we build our agents through the standard reinforcement learning algorithm in curriculum learning phases and leverage centralized control via a pixel-to-pixel policy network. As agents co-evolve through self-play, we observe several stages of intelligence, from the acquisition of atomic skills to the development of group strategies. Since these learned group strategies arise from individual decisions without an explicit coordination mechanism, we claim that artificial collective intelligence emerges from massive-agent cooperation and competition. We further analyze the emergence of various learned strategies through metrics and ablation studies, aiming to provide insights for reinforcement learning implementations in massive-agent environments.

****

Offline Reinforcement Learning via High-Fidelity Generative Behavior Modeling

Huayu Chen,Cheng Lu,Chengyang Ying,Hang Su,Jun Zhu

In offline reinforcement learning, weighted regression is a common method to ensure the learned policy stays close to the behavior policy and to prevent selecting out-of-sample actions. In this work, we show that due to the limited distributional expressivity of policy models, previous methods might still select unseen actions during training, which deviates from their initial motivation. To address this problem, we adopt a generative approach by decoupling the learned policy into two parts: an expressive generative behavior model and an action evaluation model. The key insight is that such decoupling avoids learning an explicitly parameterized policy model with a closed-form expression. Directly learning the behavior policy allows us to leverage existing advances in generative modeling, such as diffusion-based methods, to model diverse behaviors. As for action evaluation, we combine our method with an in-sample planning technique to further avoid selecting out-of-sample actions and increase computational efficiency. Experimental results on D4RL datasets show that our proposed method achieves competitive or superior performance compared with state-of-the-art offline RL methods, especially in complex tasks such as AntMaze. We also empirically demonstrate that our method can successfully learn from a heterogeneous dataset containing multiple distinctive but similarly successful strategies, whereas previous unimodal policies fail.

****

GraphVF: Controllable Protein-Specific 3D Molecule Generation with Variational Flow

Fang Sun,Zhihao Zhan,Hongyu Guo,Ming Zhang,Jian Tang

Designing molecules that bind to specific target proteins is a fundamental task in drug discovery. Recent generative models leveraging geometrical constraints imposed by proteins and molecules have shown great potential in generating protein-specific 3D molecules. Nevertheless, these existing methods fail to generate 3D molecules with 2D skeletal curtailments, which encode pharmacophoric patterns essential to drug potency. To cope with this challenge, we propose GraphVF, which seamlessly integrates geometrical and skeletal restraints into a variational flow framework, where the former is captured through a flow transformation and the latter is encoded by an amortized factorized Gaussian. We empirically verify that our method achieves state-of-the-art performance on protein-specific 3D molecule generation in terms of binding affinity and some other drug properties. In particular, it represents the first controllable geometry-aware, protein-specific molecule generation method, which enables creating 3D molecules with specified chemical sub-structures or drug properties.

****

Graph Neural Networks as Gradient Flows: understanding graph convolutions via energy

Francesco Di Giovanni,James Rowbottom,Benjamin Paul Chamberlain,Thomas Markovich,Michael M. Bronstein

Gradient flows are differential equations that minimize an energy functional and constitute the main descriptors of physical systems. We apply this formalism to Graph Neural Networks (GNNs) to develop new frameworks for learning on graphs as well as provide a better theoretical understanding of existing ones. We derive GNNs as a gradient flow equation of a parametric energy that provides a physics-inspired interpretation of GNNs as learning particle dynamics in the feature space. In particular, we show that in graph convolutional models (GCN), the positive/negative eigenvalues of the channel mixing matrix correspond to attractive/repulsive forces between adjacent features. We rigorously prove how the channel-mixing can learn to steer the dynamics towards low or high frequencies, which allows to deal with heterophilic graphs. We show that the same class of energies is decreasing along a larger family of GNNs; albeit not gradient flows, they retain their inductive bias. We experimentally evaluate an instance of the gradient flow framework that is principled, more efficient than GCN, and achieves competiti ve performance on graph datasets of varying homophily often outperforming recent baselines specifically designed to target heterophily.

****

CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers

Wenyi Hong,Ming Ding,Wendi Zheng,Xinghan Liu,Jie Tang

In this work, we present CogVideo, a 9B-parameter transformer for text-to-video generation. The CogVideo model has been trained by inheriting a pretrained text-to-image model, CogView2, which significantly reduces the training cost and alleviates the problem of scarcity and weak relevance. We also propose a multi-frame-rate training strategy for better aligning text and video clips. CogVideo achieves state-of-the-art performance in machine evaluation and outperforms publicly available models by a large margin in human evaluation. Its codes and model are also publicly available at https://github.com/THUDM/CogVideo.

****

Revisit Finetuning strategy for Few-Shot Learning to Transfer the Emdeddings

Heng Wang,Tan Yue,Xiang Ye,Zihang He,Bohan Li,Yong Li

Few-Shot Learning (FSL) aims to learn a simple and effective bias on limited novel samples. Recently, many methods have been focused on re-training a randomly initialized linear classifier to adapt it to the novel features extracted by the pre-trained feature extractor (called Linear-Probing-based methods). These methods typically assumed the pre-trained feature extractor was robust enough, i.e., finetuning was not needed, and hence the pre-trained feature extractor does not change on the novel samples. However, the unchanged pre-trained feature extractor will distort the features of novel samples because the robustness assumption may not hold, especially on the out-of-distribution samples. To extract the undistorted features, we designed Linear-Probing-Finetuning with Firth-Bias (LP-FT-FB) to yield an accurate bias on the limited samples for better finetuning the pre-trained feature extractor, providing stronger transferring ability. In LP-FT-FB, we further proposed inverse Firth Bias Reduction (i-FBR) to regularize the over-parameterized feature extractor on which FBR does not work well.■The proposed i-FBR effectively alleviates the over-fitting problem of the feature extractor in the process of finetuning and helps extract undistorted novel features. To show the effectiveness of the designed LP-FT-FB, we conducted a lot of experiments on the commonly used FSL datasets under different backbones, including in-domain and cross-domain FSL tasks. The experimental results show that the proposed FT-LP-FB outperforms the SOTA FSL methods. The code is available at https://github.com/whzyf951620/LinearProbingFinetuningFirthBias.

****

Scalable Multi-Modal Continual Meta-Learning

Bin Wu,Shangsong Liang,Qiang Zhang

This paper focuses on continual meta-learning, where few-shot tasks are sequentially available and sampled from a non-stationary distribution. Motivated by this challenging setting, many works have been developed with a mixture of meta-knowledge to cope with the heterogeneity and a dynamically changing number of components to capture incremental information. However, the underlying assumption of mutual exclusiveness among mixture components prevents sharing meta-knowledge across different clusters of tasks. Moreover, the existing incremental methods only rely on the prior to determine whether to increase meta-knowledge, where the unlimited increase would lead to parameter inefficiency. In our work, we propose a Scalable Multi-Modal Continual Meta-Learning (SMM-CML) algorithm. It employs a multi-modal premise that not only encourages different clusters of tasks to share meta-knowledge but also maintains their diversity. Moreover, to capture the incremental information, our algorithm uses Indian Buffet Process (IBP) as a prior number of components and proposes a sparsity method based on evidential theory to filter out the components without receiving support information directly from tasks. Thus we can learn the posterior number of components to avoid parameter inefficiency and reduce computational consumption. Experiments show SMM-CML outperforms SOTA baselines, which illustrates the effectiveness of our multi-modal meta-knowledge, and confirms that our algorithm can learn the really need meta-knowledge from tasks.

****

Optimizing Spca-based Continual Learning: A Theoretical Approach

Chunchun Yang,Malik Tiomoko,Zengfu Wang

Catastrophic forgetting and the stability-plasticity dilemma are two major obstacles to continual learning. In this paper we first propose a theoretical analysis of a SPCA-based continual learning algorithm using high dimensional statistics. Second, we design OSCL (Optimized Spca-based Continual Learning) which builds on a flexible task optimization based on the theory. By optimizing a single task, catastrophic forgetting can be prevented theoretically. While optimizing multi-tasks, the trade-off between integrating knowledge from the new task and retaining previous knowledge of the old task can be achieved by assigning appropriate weights to corresponding tasks in compliance with the objectives. Experimental results confirm that the various theoretical conclusions are robust to a wide range of data distributions. Besides, several applications on synthetic and real data show that the proposed method while being computationally efficient, achieves comparable results with some state of the art.

****

Value Memory Graph: A Graph-Structured World Model for Offline Reinforcement Learning

Deyao Zhu,Li Erran Li,Mohamed Elhoseiny

Reinforcement Learning (RL) methods are typically applied directly in environments to learn policies. In some complex environments with continuous state-action spaces, sparse rewards, and/or long temporal horizons, learning a good policy in the original environments can be difficult. Focusing on the offline RL setting, we aim to build a simple and discrete world model that abstracts the original environment. RL methods are applied to our world model instead of the environment data for simplified policy learning. Our world model, dubbed Value Memory Graph (VMG), is designed as a directed-graph-based Markov decision process (MDP) of which vertices and directed edges represent graph states and graph actions, separately. As state-action spaces of VMG are finite and relatively small compared to the original environment, we can directly apply the value iteration algorithm on VMG to estimate graph state values and figure out the best graph actions. VMG is trained from and built on the offline RL dataset. Together with an action translator that converts the abstract graph actions in VMG to real actions in the original environment, VMG controls agents to maximize episode returns. Our experiments on the D4RL benchmark show that VMG can outperform state-of-the-art offline RL methods in several tasks, especially when environments have sparse rewards and long temporal horizons. Code is available at https://github.com/TsuTikgiau/ValueMemoryGraph

****

CAKE: CAusal and collaborative proxy-tasKs lEarning for Semi-Supervised Domain Adaptation

Wenqiao Zhang,CHANGSHUO LIU,Can Cui,Beng Chin Ooi

Semi-supervised domain adaptation (SSDA) adapts a learner to a new domain by effectively utilizing source domain data and a few labeled target samples. It is a practical yet under-investigated research topic. In this paper, we analyze the SSDA problem from two perspectives that have previously been overlooked, and correspondingly decompose it into two \emph{key subproblems}: \emph{robust domain adaptation (DA) learning} and \emph{maximal cross-domain data utilization}. \textbf{(i)} From a causal theoretical view, a robust DA model should distinguish the invariant ``concept'' (key clue to image label) from the nuisance of confounding factors across domains. To achieve this goal, we propose to generate \emph{concept-invariant samples} to enable the model to classify the samples through causal intervention, yielding improved generalization guarantees; \textbf{(ii)} Based on the robust DA theory, we aim to exploit the maximal utilization of rich source domain data and a few labeled target samples to boost SSDA further. Consequently, we propose a collaboratively debiasing learning framework that utilizes two complementary semi-supervised learning (SSL) classifiers to mutually exchange their unbiased knowledge, which helps unleash the potential of source and target domain training data, thereby producing more convincing pseudo-labels. Such ob tained labels facilitate cross-domain feature alignment and duly improve the invariant concept learning. In our experimental study, we show that the proposed model significantly outperforms SOTA methods in terms of effectiveness and generalisability on SSDA datasets.

****

RulE: Neural-Symbolic Knowledge Graph Reasoning with Rule Embedding
Xiaojuan Tang,Song-Chun Zhu,Yitao Liang,Muhan Zhang

Knowledge graph (KG) reasoning is an important problem for knowledge graphs. It predicts missing links by reasoning on existing facts. Knowledge graph embedding (KGE) is one of the most popular methods to address this problem. It embeds entities and relations into low-dimensional vectors and uses the learned entity/relation embeddings to predict missing facts. However, KGE only uses zeroth-order (propositional) logic to encode existing triplets (e.g., ``Alice is Bob's wife."); it is unable to leverage first-order (predicate) logic to represent generally applicable logical \textbf{rules} (e.g., ``$\forall x,y \colon x ~\text{is}~ y\text{'s wife} \rightarrow y ~\text{is}~ x\text{'s husband}$''). On the other hand, traditional rule-based KG reasoning methods usually rely on hard logical rule inference, making it brittle and hardly competitive with KGE. In this paper, we propose RulE, a novel and principled framework to represent and model logical rules and triplets. RulE jointly represents entities, relations and logical rules in a unified embedding space. By learning an embedding for each logical rule, RulE can perform logical rule inference in a soft way and give a confidence score to each grounded rule, similar to how KGE gives each triplet a confidence score. Compared to KGE alone, RulE allows injecting prior logical rule information into the embedding space, which improves the generalization of knowledge graph embedding. Besides, the learned confidence scores of rules improve the logical rule inference process by softly controlling the contribution of each rule, which alleviates the brittleness of logic. We evaluate our method with link prediction tasks. Experimental results on multiple benchmark KGs demonstrate the effectiveness of RulE.

****

Sampling-free Inference for Ab-Initio Potential Energy Surface Networks
Nicholas Gao,Stephan Günnemann

Recently, it has been shown that neural networks not only approximate the ground-state wave functions of a single molecular system well but can also generalize to multiple geometries. While such generalization significantly speeds up training, each energy evaluation still requires Monte Carlo integration which limits the evaluation to a few geometries. In this work, we address the inference shortcomings by proposing the Potential learning from ab-initio Networks (PlaNet) framework, in which we simultaneously train a surrogate model in addition to the neural wave function. At inference time, the surrogate avoids expensive Monte-Carlo integration by directly estimating the energy, accelerating the process from hours to milliseconds. In this way, we can accurately model high-resolution multi-dimensional energy surfaces for larger systems that previously were unobtainable via neural wave functions. Finally, we explore an additional inductive bias by introducing physically-motivated restricted neural wave function models. We implement such a function with several additional improvements in the new PESNet++ model. In our experimental evaluation, PlaNet accelerates inference by 7 orders of magnitude for larger molecules like ethanol while preserving accuracy. Compared to previous energy surface networks, PESNet++ reduces energy errors by up to 74%.

****

Hidden Schema Networks
Ramses J Sanchez,Lukas Alexander Conrads,Pascal Welke,Kostadin Cvejoski,Cesar Ojeda

Most modern language models infer representations that, albeit powerful, lack both compositionality and semantic interpretability. Starting from the assumption that a large proportion of semantic content is necessarily relational, we introduce a neural language model that discovers networks of symbols (schemata) from text datasets. Using a variational autoencoder (VAE) framework, our model encodes sentences into sequences of symbols (composed representation), which correspond to the nodes visited by biased random walkers on a global latent graph. We first demonstrate that the model is able to uncover ground-truth graphs from artificially generated datasets of random token sequences. Next we leverage pretrained BERT and GPT-2 language models as encoder and decoder, respectively, to train our model on language modelling and commonsense knowledge generation tasks. Qualitatively, the model is able to infer schema networks whose nodes (symbols) can be interpreted as encoding different aspects of natural language (as e.g. topics, sentiments). Quantitatively, our results show that the model successfully interprets the encoded symbol sequences, as it achieves state-of-the-art scores on VAE language modeling benchmarks. Source code to reproduce all experiments is provided with the supplementary material.

****

$\mathscr{N}$-WL: A New Hierarchy of Expressivity for Graph Neural Networks

Qing Wang,Dillon Ze Chen,Asiri Wijesinghe,Shouheng Li,Muhammad Farhan

The expressive power of Graph Neural Networks (GNNs) is fundamental for understanding their capabilities and limitations, i.e., what graph properties can or cannot be learnt by a GNN. Since standard GNNs have been characterised to be upper-bounded by the Weisfeiler-Lehman (1-WL) algorithm, recent attempts concentrated on developing more expressive GNNs in terms of the k-WL hierarchy, a well-established framework for graph isormorphism tests. In this work we show that, contrary to the widely accepted view, the k-WL hierarchy is not well-suited for measuring expressive GNNs. This is due to limitations that are inherent to high-dimensional WL algorithms such as the lack of a natural interpretation and high computational costs, which makes it difficult to draw any firm conclusions about the expressive power of GNNs beyond 1-WL. Thus, we propose a novel hierarchy of graph isomorphism tests, namely Neighbourhood WL ($\mathscr{N}$-WL), and also establish a new theorem on the equivalence of expressivity between induced connected subgraphs and induced subgraphs within this hierarchy. Further, we design a GNN model upon $\mathscr{N}$-WL, Graph Neighbourhood Neural Network (G3N), and empirically verify its expressive power on synthetic and real-world benchmarks.

****

Learning Input-agnostic Manipulation Directions in StyleGAN with Text Guidance

Yoonjeon Kim,Hyunsu Kim,Junho Kim,Yunjey Choi,Eunho Yang

With the advantages of fast inference and human-friendly flexible manipulation, image-agnostic style manipulation via text guidance enables new applications that were not previously available. The state-of-the-art text-guided image-agnostic manipulation method embeds the representation of each channel of StyleGAN independently in the Contrastive Language-Image Pre-training (CLIP) space, and provides it in the form of a Dictionary to quickly find out the channel-wise manipulation direction during inference time. However, in this paper we argue that this dictionary which is constructed by controlling single channel individually is limited to accommodate the versatility of text guidance since the collective and interactive relation among multiple channels are not considered. Indeed, we show that it fails to discover a large portion of manipulation directions that can be found by existing methods, which manually manipulates latent space without texts. To alleviate this issue, we propose a novel method that learns a Dictionary, whose entry corresponds to the representation of a single channel, by taking into account the manipulation effect coming from the interaction with multiple other channels. We demonstrate that our strategy resolves the inability of previous methods in finding diverse known directions from unsupervised methods and unknown directions from random text while maintaining the real-time inference speed and disentanglement ability.

****

End-to-end Invariance Learning with Relational Inductive Biases in Multi-Object Robotic Manipulation

Davide Mambelli,Frederik Träuble,Stefan Bauer,Bernhard Schölkopf,Francesco Locatello

Although reinforcement learning has seen remarkable progress over the last years, solving robust dexterous object-manipulation tasks in multi-object settings re mains a challenge. In this paper, we focus on models that can learn manipulation tasks in fixed multi-object settings \emph{and} extrapolate this skill zero-shot without any drop in performance when the number of objects changes. We consider the generic task of moving a single cube out of a set to a goal position. We find that previous approaches, which primarily leverage attention and graph neural network-based architectures, do not exhibit this invariance when the number of input objects changes while scaling as K^2. We analyse effects on generalization of different relational inductive biases and then propose an efficient plug-and-play module that overcomes these limitations. Besides exceeding performances in their training environment, we show that our approach, which scales linearly in K, allows agents to extrapolate and generalize zero-shot to any new object number.

****

DAVA: Disentangling Adversarial Variational Autoencoder

Benjamin Estermann,Roger Wattenhofer

The use of well-disentangled representations offers many advantages for downstream tasks, e.g. an increased sample efficiency, or better interpretability.

However, the quality of disentangled interpretations is often highly dependent on the choice of dataset-specific hyperparameters, in particular the regularization strength.

To address this issue, we introduce DAVA, a novel training procedure for variational auto-encoders. DAVA completely alleviates the problem of hyperparameter selection.

We compare DAVA to models with optimal hyperparameters.

Without any hyperparameter tuning, DAVA is competitive on a diverse range of commonly used datasets.

Underlying DAVA, we discover a necessary condition for unsupervised disentanglement, which we call PIPE.

We demonstrate the ability of PIPE to positively predict the performance of downstream models in abstract reasoning.

We also thoroughly investigate correlations with existing supervised and unsupervised metrics. The code is available at https://github.com/besterma/dava.

****

TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations

Haoxuan Li,Yan Lyu,Chunyuan Zheng,Peng Wu

Bias is a common problem inherent in recommender systems, which is entangled with users' preferences and poses a great challenge to unbiased learning. For debiasing tasks, the doubly robust (DR) method and its variants show superior performance due to the double robustness property, that is, DR is unbiased when either imputed errors or learned propensities are accurate.

However, our theoretical analysis reveals that DR usually has a large variance. Meanwhile, DR would suffer unexpectedly large bias and poor generalization caused by inaccurate imputed errors and learned propensities, which usually occur in practice. In this paper, we propose a principled approach that can effectively reduce the bias and variance simultaneously for existing DR approaches when the error imputation model is misspecified. In addition, we further propose a novel semi-parametric collaborative learning approach that decomposes imputed errors into parametric and nonparametric parts and updates them collaboratively, resulting in more accurate predictions. Both theoretical analysis and experiments demonstrate the superiority of the proposed methods compared with existing debiasing methods.

****

DeepGRAND: Deep Graph Neural Diffusion

Khang Nguyen,Nong Minh Hieu,Tan Minh Nguyen,Nguyen Duy Khuong,Vinh Duc NGUYEN

We propose the Deep Graph Neural Diffusion (DeepGRAND), a class of continuous-depth graph neural networks based on the diffusion process on graphs. DeepGRAND leverages a data-dependent scaling term and a perturbation to the graph diffusivity to make the real part of all eigenvalues of the diffusivity matrix become negative, which ensures two favorable theoretical properties: (i) the node represent ation does not exponentially converge to a constant vector as the model depth increases, thus alleviating the over-smoothing issue; (ii) the stability of the model is guaranteed by controlling the norm of the node representation. Compared to the baseline GRAND, DeepGRAND mitigates the accuracy drop-off with increasing depth and improves the overall accuracy of the model. We empirically corroborate the advantage of DeepGRAND over many existing graph neural networks on various graph deep learning benchmark tasks.

****

Learning Discrete Representation with Optimal Transport Quantized Autoencoders

Xiaoyu Bie,Dexiong Chen,Xiaodong Cun,Xi SHEN

Vector quantized variational autoencoder (VQ-VAE) has recently emerged as a powerful generative model for learning discrete representations. Like other vector quantization methods, one key challenge of training VQ-VAE comes from the codebook collapse, i.e. only a fraction of codes are used, limiting its reconstruction qualities. To this end, VQ-VAE often leverages some carefully designed heuristics during the training to use more codes. In this paper, we propose a simple yet effective approach to overcome this issue through optimal transport, which regularizes the quantization by explicitly assigning equal number of samples to each code. The proposed approach, named OT-VAE, enforces the full utilization of the codebook while not requiring any heuristics. We empirically validate our approach on three different data modalities: images, speech, and 3D human motions. For all the modalities, OT-VAE shows better reconstruction with higher perplexity than other VQ-VAE variants on several datasets. In particular, OT-VAE achieves state-of-the-art results on the AIST++ dataset for 3D dance generation. Our code will be released upon publication.

****

How to Keep Cool While Training

Martin Trimmel,Mihai Zanfir,Richard Hartley,Cristian Sminchisescu

Modern classification neural networks are notoriously prone to being overly confident in their predictions. With multiple calibration methods having been proposed so far, there has been noteworthy progress in reducing this overconfidence. However, to the best of our knowledge, prior methods have exclusively focused on the factors that affect calibration, leaving open the reverse question of how (mis)calibration impacts network training. Aiming for a better understanding of this interplay, we propose a temperature-based Cooling method for calibrating classification neural networks during training. Cooling has a substantial effect on the gradients and reduces the need for a learning rate schedule. We investigate different variants of Cooling, with the simplest one, last layer Cooling, being also the best-performant one, improving network performance on a range of datasets, network architectures, and hyperparameter settings.

****

Learning System Dynamics from Sensory Input under Optimal Control Principles

Oumayma Bounou,Jean Ponce,Justin Carpentier

Identifying the underlying dynamics of actuated physical systems from sensory input is of high interest in control, robotics, and engineering in general. In the context of control problems, existing approaches decouple the construction of the feature space where the dynamics identification process occurs from the target control tasks, potentially leading to a mismatch between feature and real state spaces: the systems may not be controllable in feature space, and synthesized controls may not be applicable in the state space. Borrowing from the Koopman formalism, we propose instead to learn an embedding of both the states and con- trols in feature spaces where the dynamics are linear, and to include the target control task in the learning objective in the form of a differentiable and robust optimal control problem. We validate this approach with simulation experiments of systems with non-linear dynamics, demonstrating that the controls obtained in feature space can be used to drive the corresponding physical systems and that the learned model can serve for future state prediction.

****

Dual Algorithmic Reasoning

Danilo Numeroso,Davide Bacciu,Petar Veli■kovi■

Neural Algorithmic Reasoning is an emerging area of machine learning which seeks to infuse algorithmic computation in neural networks, typically by training neural models to approximate steps of classical algorithms. In this context, much of the current work has focused on learning reachability and shortest path graph algorithms, showing that joint learning on similar algorithms is beneficial for generalisation. However, when targeting more complex problems, such "similar" algorithms become more difficult to find. Here, we propose to learn algorithms by exploiting duality of the underlying algorithmic problem. Many algorithms solve optimisation problems. We demonstrate that simultaneously learning the dual definition of these optimisation problems in algorithmic learning allows for better learning and qualitatively better solutions. Specifically, we exploit the max-flow min-cut theorem to simultaneously learn these two algorithms over synthetically generated graphs, demonstrating the effectiveness of the proposed approach. We then validate the real-world utility of our dual algorithmic reasoner by deploying it on a challenging brain vessel classification task, which likely depends on the vessels’ flow properties. We demonstrate a clear performance gain when using our model within such a context, and empirically show that learning the max-flow and min-cut algorithms together is critical for achieving such a result.

****

Lmser-pix2seq: Learning Stable Sketch Representations For Sketch Healing
Tengjie Li,Sicong Zang,Shikui Tu,Lei Xu

Sketch healing aims to recreate a complete sketch from the corrupted one. The sparse and abstract nature of the sketch makes it challenging due to the difficulty in learning. The features extracted from the corrupted sketch may be inconsistent with the ones from the corresponding full sketch. In this paper, we present Lmser-pix2seq to learn stable sketch representations against the missing information by employing a Least mean square error reconstruction (Lmser) block, which falls into encoder-decoder paradigm. Taking as input a corrupted sketch, the Lmser encoder computes the embeddings of structural patterns of the input, while the decoder reconstructs the complete sketch from the embeddings. We build bi-directional skip connections between the encoder and the decoder in our Lmser block. The feedback connections enable recurrent paths to receive more information about the reconstructed sketch produced by the decoder, which helps the encoder extract stable sketch features. The features captured by the Lmser block are eventually fed into a recurrent neural network decoder to recreate the sketches. Experimental results show that our Lmser-pix2seq outperforms the state-of-the-art methods in sketch healing, especially when the sketches are heavily masked or corrupted.

****

Toward Effective Deep Reinforcement Learning for 3D Robotic Manipulation: End-to-End Learning from Multimodal Raw Sensory Data
Samyeul Noh,Hyun Myung

Sample-efficient reinforcement learning (RL) methods capable of learning directly from raw sensory data without the use of human-crafted representations would open up real-world applications in robotics and control. Recent advances in visual RL have shown that learning a latent representation together with existing RL algorithms closes the gap between state-based and image-based training. However, image-based training is still significantly sample-inefficient with respect to learning in 3D continuous control problems (for example, robotic manipulation) compared to state-based training. In this study, we propose an effective model-free off-policy RL method for 3D robotic manipulation that can be trained in an end-to-end manner from multimodal raw sensory data obtained from a vision camera and a robot's joint encoders, without the need for human-crafted representations. Notably, our method is capable of learning a latent multimodal representation and a policy in an efficient, joint, and end-to-end manner from multimodal raw sensory data. Our method, which we dub MERL: Multimodal End-to-end Reinforcement Learning, results in a simple but effective approach capable of significantly outperforming both current state-of-the-art visual RL and state-based RL methods with respect to sample efficiency, learning performance, and training stability in relation to 3D robotic manipulation tasks from DeepMind Control.

****

Domain Generalisation via Domain Adaptation: An Adversarial Fourier Amplitude Approach

Minyoung Kim,Da Li,Timothy Hospedales

We tackle the domain generalisation (DG) problem by posing it as a domain adaptation (DA) task where we adversarially synthesise the worst-case `target' domain and adapt a model to that worst-case domain, thereby improving the model’s robustness. To synthesise data that is challenging yet semantics-preserving, we generate Fourier amplitude images and combine them with source domain phase images, exploiting the widely believed conjecture from signal processing that amplitude spectra mainly determines image style, while phase data mainly captures image semantics. To synthesise a worst-case domain for adaptation, we train the classifier and the amplitude generator adversarially. Specifically, we exploit the maximum classifier discrepancy (MCD) principle from DA that relates the target domain performance to the discrepancy of classifiers in the model hypothesis space. By Bayesian hypothesis modeling, we express the model hypothesis space effectively as a posterior distribution over classifiers given the source domains, making adversarial MCD minimisation feasible. On the DomainBed benchmark including the large-scale DomainNet dataset, the proposed approach yields significantly improved domain generalisation performance over the state-of-the-art.

****

Improving Generative Flow Networks with Path Regularization

Anh Do,Duy Dinh,Tan Minh Nguyen,Nguyen Duy Khuong,Stanley Osher,Nhat Ho

Generative Flow Networks (GFlowNets) are recently proposed models for learning stochastic policies that generate compositional objects by sequences of actions with the probability proportional to a given reward function. The central problem of GFlowNets is to improve their exploration and generalization. In this work, we propose a novel path regularization method based on optimal transport theory that places prior constraints on the underlying structure of the GFlowNets. The prior is designed to help the GFlowNets better discover the latent structure of the target distribution or enhance its ability to explore the environment in the context of active learning. The path regularization controls the flow in GFlowNets to generate more diverse and novel candidates via maximizing the optimal transport distances between two forward policies or to improve the generalization via minimizing the optimal transport distances. In addition, we derive an efficient implementation of the regularization by finding its closed form solutions in specific cases and a meaningful upper bound that can be used as an approximation to minimize the regularization term. We empirically demonstrate the advantage of our path regularization on a wide range of tasks, including synthetic hypergrid environment modeling, discrete probabilistic modeling, and biological sequence design.

****

Confidential-PROFITT: Confidential PROof of FaIr Training of Trees

Ali Shahin Shamsabadi,Sierra Calanda Wyllie,Nicholas Franzese,Natalie Dullerud,Sébastien Gambs,Nicolas Papernot,Xiao Wang,Adrian Weller

Post hoc auditing of model fairness suffers from potential drawbacks: (1) auditing may be highly sensitive to the test samples chosen; (2) the model and/or its training data may need to be shared with an auditor thereby breaking confidentiality. We address these issues by instead providing a certificate that demonstrates that the learning algorithm itself is fair, and hence, as a consequence, so too is the trained model. We introduce a method to provide a confidential proof of fairness for training, in the context of widely used decision trees, which we term Confidential-PROFITT. We propose novel fair decision tree learning algorithms along with customized zero-knowledge proof protocols to obtain a proof of fairness that can be audited by a third party. Using zero-knowledge proofs enables us to guarantee confidentiality of both the model and its training data. We show empirically that bounding the information gain of each node with respect to the sensitive attributes reduces the unfairness of the final tree. In extensive experiments on the COMPAS, Communities and Crime, Default Credit, and Adult datasets, we demonstrate that a company can use Confidential-PROFITT to certify the fai rness of their decision tree to an auditor in less than 2 minutes, thus indicating the applicability of our approach. This is true for both the demographic parity and equalized odds definitions of fairness. Finally, we extend Confidential-PROFITT to apply to ensembles of trees.

****

Consolidator: Mergable Adapter with Group Connections for Visual Adaptation

Tianxiang Hao,Hui Chen,Yuchen Guo,Guiguang Ding

Recently, transformers have shown strong ability as visual feature extractors, surpassing traditional convolution-based models in various scenarios. However, the success of vision transformers largely owes to their capacity to accommodate numerous parameters. As a result, new challenges for adapting a well-trained transformer to downstream tasks arise. On the one hand, classic fine-tuning tunes all parameters in a huge model for every downstream task and thus easily falls into an overfitting situation, leading to inferior performance. On the other hand, on resource-limited devices, fine-tuning stores a full copy of all parameters and thus is usually impracticable for the shortage of storage space. However, few works have focused on how to efficiently and effectively transfer knowledge in a vision transformer. Existing methods did not dive into the properties of visual features, leading to inferior performance. Moreover, some of them bring heavy inference cost though benefiting storage. To tackle these problems, we propose consolidator to achieve efficient transfer learning for large vision models. Our consolidator modifies the pre-trained model with the addition of a small set of tunable parameters to temporarily store the task-specific knowledge while freezing the backbone model during adaptation. Motivated by the success of group-wise convolution, we adopt grouped connections across the features extracted by fully connected layers to construct tunable parts in a consolidator. To further enhance the model's capacity to transfer knowledge under a constrained storage budget and keep inference efficient, we consolidate the parameters in two stages: 1. between adaptation and storage, and 2. between loading and inference. On a series of downstream visual tasks, our consolidator can reach up to 7.56 better accuracy than full fine-tuning with merely 0.35% parameters, and outperform state-of-the-art parameter-efficient tuning methods by a clear margin. Code is available at github.

****

Statistical Theory of Differentially Private Marginal-based Data Synthesis Algorithms

Ximing Li,Chendi Wang,Guang Cheng

Marginal-based methods achieve promising performance in the synthetic data competition hosted by the National Institute of Standards and Technology (NIST).

To deal with high-dimensional data, the distribution of synthetic data is represented by a probabilistic graphical model (e.g., a Bayesian network), while the raw data distribution is approximated by a collection of low-dimensional marginals.

Differential privacy (DP) is guaranteed by introducing random noise to each low-dimensional marginal distribution.

Despite its promising performance in practice, the statistical properties of marginal-based methods are rarely studied in the literature.

In this paper, we study DP data synthesis algorithms based on Bayesian networks (BN) from a statistical perspective. We establish a rigorous accuracy guarantee for BN-based algorithms, where the errors are measured by the total variation (TV) distance or the L^2 distance.

Related to downstream machine learning tasks, an upper bound for the utility error of the DP synthetic data is also derived. To complete the picture, we establish a lower bound for TV accuracy that holds for every ϵ-DP synthetic data generator.

****

Homotopy-based training of NeuralODEs for accurate dynamics discovery

Joon-Hyuk Ko,Hankyul Koh,Nojun Park,Wonho Jhe

Conceptually, Neural Ordinary Differential Equations (NeuralODEs) pose an attractive way to extract dynamical laws from time series data, as they are natural ex tensions of the traditional differential equation-based modeling paradigm of the physical sciences. In practice, NeuralODEs display long training times and suboptimal results, especially for longer duration data where they may fail to fit the data altogether. While methods have been proposed to stabilize NeuralODE training, many of these involve placing a strong constraint on the functional form the trained NeuralODE can take that the actual underlying governing equation does not guarantee satisfaction. In this work, we present a novel NeuralODE training algorithm that leverages tools from the chaos and mathematical optimization communities -- synchronization and homotopy optimization -- for a breakthrough in tackling the NeuralODE training obstacle. We demonstrate architectural changes are unnecessary for effective NeuralODE training. Compared to the conventional training methods, our algorithm achieves drastically lower loss values without any changes to the model architectures. Experiments on both simulated and real systems with complex temporal behaviors demonstrate NeuralODEs trained with our algorithm are able to accurately capture true long term behaviors and correctly extrapolate into the future.

****

Transformers with Multiresolution Attention Heads
Tan Minh Nguyen,Tho Tran Huu,Tam Minh Nguyen,Minh Pham,Nhat Ho,Stanley Osher

We propose the Transformer with Multiresolution-head Attention (MrsFormer), a class of efficient transformers inspired by the multiresolution approximation (MRA) for approximating a signal f using wavelet bases. MRA decomposes a signal into components that lie on orthogonal subspaces at different scales. Similarly, MrsFormer decomposes the attention heads in the multi-head attention into fine-scale and coarse-scale heads, modeling the attention patterns between tokens and between groups of tokens. Computing the attention heads in MrsFormer requires significantly less computation and memory footprint compared to the standard softmax transformer with multi-head attention. We analyze and validate the advantage of MrsFormer over the standard transformers on a wide range of applications including image and time series classification.

****

Anti-Symmetric DGN: a stable architecture for Deep Graph Networks
Alessio Gravina,Davide Bacciu,Claudio Gallicchio

Deep Graph Networks (DGNs) currently dominate the research landscape of learning from graphs, due to their efficiency and ability to implement an adaptive message-passing scheme between the nodes. However, DGNs are typically limited in their ability to propagate and preserve long-term dependencies between nodes, i.e., they suffer from the over-squashing phenomena. As a result,
we can expect them to under-perform, since different problems require to capture interactions at different (and possibly large) radii in order to be effectively solved. In this work, we present Anti-Symmetric Deep Graph Networks (A-DGNs), a framework for stable and non-dissipative DGN design, conceived through the lens of ordinary differential equations. We give theoretical proof that our method is stable and non-dissipative, leading to two key results: long-range information between nodes is preserved, and no gradient vanishing or explosion occurs in training. We empirically validate the proposed approach on several graph benchmarks, showing that A-DGN yields to improved performance and enables to learn effectively even when dozens of layers are used.

****

Contrastive Learning for Unsupervised Domain Adaptation of Time Series
Yilmazcan Ozyurt,Stefan Feuerriegel,Ce Zhang

Unsupervised domain adaptation (UDA) aims at learning a machine learning model using a labeled source domain that performs well on a similar yet different, unlabeled target domain. UDA is important in many applications such as medicine, where it is used to adapt risk scores across different patient cohorts. In this paper, we develop a novel framework for UDA of time series data, called CLUDA. Specifically, we propose a contrastive learning framework to learn contextual representations in multivariate time series, so that these preserve label information for the prediction task. In our framework, we further capture the variation in t he contextual representations between source and target domain via a custom nearest-neighbor contrastive learning. To the best of our knowledge, ours is the first framework to learn domain-invariant, contextual representation for UDA of time series data. We evaluate our framework using a wide range of time series datasets to demonstrate its effectiveness and show that it achieves state-of-the-art performance for time series UDA.

****

Model-Based Decentralized Policy Optimization

Hao Luo,Jiechuan Jiang,Zongqing Lu

Decentralized policy optimization has been commonly used in cooperative multi-agent tasks. However, since all agents are updating their policies simultaneously, from the perspective of individual agents, the environment is non-stationary, resulting in it being hard to guarantee monotonic policy improvement. To help the policy improvement be stable and monotonic, we propose model-based decentralized policy optimization (MDPO), which incorporates a latent variable function to help construct the transition and reward function from an individual perspective. We theoretically analyze that the policy optimization of MDPO is more stable than model-free decentralized policy optimization. Moreover, due to non-stationarity, the latent variable function is varying and hard to be modeled. We further propose a latent variable prediction method to reduce the error of latent variable function, which theoretically contributes to the monotonic policy improvement. Empirically, MDPO can indeed obtain superior performance than model-free decentralized policy optimization in a variety of cooperative multi-agent tasks.

****

Online Low Rank Matrix Completion

Soumyabrata Pal,Prateek Jain

We study the problem of online low-rank matrix completion with M users, N items and T rounds. In each round, the algorithm recommends one item per user, for which it gets a (noisy) reward sampled from a low-rank user-item preference matrix. The goal is to design a method with sub-linear regret (in T) and nearly optimal dependence on M and N. The problem can be easily mapped to the standard multi-armed bandit problem where each item is an independent arm, but that leads to poor regret as the correlation between arms and users is not exploited. On the other hand, exploiting the low-rank structure of reward matrix is challenging due to non-convexity of the low-rank manifold. We first demonstrate that the low-rank structure can be exploited using a simple explore-then-commit (ETC) approach that ensures a regret of $O(\mathsf{polylog} (\mathsf{M}+\mathsf{N}) \mathsf{T}^{2/3})$. That is, roughly only $\mathsf{polylog} (\mathsf{M}+\mathsf{N})$ item recommendations are required per user to get a non-trivial solution. We then improve our result for the rank-1 setting which in itself is quite challenging and encapsulates some of the key issues. Here, we propose OCTAL (Online Collaborative filTering using iterAtive user cLustering) that guarantees nearly optimal regret of $O(\mathsf{polylog} (\mathsf{M}+\mathsf{N}) \mathsf{T}^{1/2})$. OCTAL is based on a novel technique of clustering users that allows iterative elimination of items and leads to a nearly optimal minimax rate.

****

Modality Complementariness: Towards Understanding Multi-modal Robustness

Siting Li,Chenzhuang Du,Yu Huang,Longbo Huang,Hang Zhao

Along with the success of multi-modal learning, the robustness of multi-modal learning is receiving attention due to real-world safety concerns. Multi-modal models are anticipated to be more robust due to the possible redundancy between modalities. However, some empirical results have offered contradictory conclusions. In this paper, we point out an essential factor that causes this discrepancy: The difference in the amount of modality-wise complementary information. We provide an information-theoretical analysis of how the modality complementariness affects the multi-modal robustness. Based on the analysis, we design a metric for quantifying how complementary the modalities are to others and propose an effective pipeline to calculate our metric. Experiments on carefully-designed synthetic data verify our theory. Further, we apply our metric to real-world multi-modal datasets and reveal their property. To our best knowledge, we are the first to identify modality complementariness as an important factor affecting multi-modal robustness.

****

Effective Offline Reinforcement Learning via Conservative State Value Estimation

Liting Chen,Jie Yan,Lu Wang,Qingwei Lin,Dongmei Zhang,Saravan Rajmohan,Thomas Moscibroda

Offline RL promises to learn effective policies from static experience datasets without further interaction, which expect to perform well in the online environment. However, it faces up to a major challenge of value over-estimation introduced by the distributional drift between the dataset and the current learned policy, which leads to learning failure in practice. The common approach is to add a penalty term to reward or value estimation in the Bellman iterations, which has given rise to a number of successful algorithms such as CQL. Meanwhile, to avoid extrapolation on unseen states, existing methods focus on conservative Q-function estimation. In this paper, we propose CSVE, a new approach that directly imposes penalty on out-of-distribution states. We prove that for the evaluated policy, our conservative state value estimation satisfies: (1) over the state distribution that samples penalizing states, it lower bounds the true values in expectation, and (2) over the marginal state distribution of data, it is no more than the true values in expectation plus a constant decided by sampling error. Further, we develop a practical actor-critic algorithm in which the critic does the conservative value estimation by additionally sampling and penalizing the states 'around' the dataset, while the actor applies advantage weighted updates to improve the policy. We evaluate in classic continual control tasks of D4RL, showing that our method performs better than the conservative Q-function learning methods (e.g., CQL) and is strongly competitive among recent SOTA methods.

****

ChemAlgebra : Algebraic Reasoning on Chemical Reactions

Andrea Valenti,Davide Bacciu,Antonio Vergari

While showing impressive performance on various kinds of learning tasks, it is yet unclear whether deep learning models have the ability to robustly tackle reasoning tasks. Measuring the robustness of reasoning in machine learning models is challenging as one needs to provide a task that cannot be easily shortcut by exploiting spurious statistical correlations in the data, while operating on complex objects and constraints. To address this issue, we propose ChemAlgebra, a benchmark for measuring the reasoning capabilities of deep learning models through the prediction of stoichiometrically-balanced chemical reactions. ChemAlgebra requires manipulating sets of complex discrete objects – molecules represented as formulas or graphs – under algebraic constraints such as the mass preservation principle. We believe that ChemAlgebra can serve as a useful test bed for the next generation of machine reasoning models and as a promoter of their development.

****

A Primal-Dual Framework for Transformers and Neural Networks

Tan Minh Nguyen,Tam Minh Nguyen,Nhat Ho,Andrea L. Bertozzi,Richard Baraniuk,Stanley Osher

Self-attention is key to the remarkable success of transformers in sequence modeling tasks including many applications in natural language processing and computer vision. Like neural network layers, these attention mechanisms are often developed by heuristics and experience. To provide a principled framework for constructing attention layers in transformers, we show that the self-attention corresponds to the support vector expansion derived from a support vector regression problem, whose primal formulation has the form of a neural network layer. Using our framework, we derive popular attention layers used in practice and propose two new attentions: 1) the Batch Normalized Attention (Attention-BN) derived from the batch normalization layer and 2) the Attention with Scaled Head (Attention-SH) derived from using less training data to fit the SVR model. We empirically demonstrate the advantages of the Attention-BN and Attention-SH in reducing head redundancy, increasing the model's accuracy, and improving the model's efficiency in a variety of practical applications including image and time-series classifi

cation.
****

Explaining RL Decisions with Trajectories
Shripad Vilasrao Deshmukh,Arpan Dasgupta,Balaji Krishnamurthy,Nan Jiang,Chirag Agarwal,Georgios Theocharous,Jayakumar Subramanian

Explanation is a key component for the adoption of reinforcement learning (RL) in many real-world decision-making problems. In the literature, the explanation is often provided by saliency attribution to the features of the RL agent's state. In this work, we propose a complementary approach to these explanations, particularly for offline RL, where we attribute the policy decisions of a trained RL agent to the trajectories encountered by it during training. To do so, we encode trajectories in offline training data individually as well as collectively (encoding a set of trajectories). We then attribute policy decisions to a set of trajectories in this encoded space by estimating the sensitivity of the decision with respect to that set. Further, we demonstrate the effectiveness of the proposed approach in terms of quality of attributions as well as practical scalability in diverse environments that involve both discrete and continuous state and action spaces such as grid-worlds, video games (Atari) and continuous control (MuJoCo). We also conduct a human study on a simple navigation task to observe how their understanding of the task compares with data attributed for a trained RL policy.

****

Keypoint Matching via Random Network Consensus
Siyan Dong,Shuzhe Wang,Daniel Barath,Juho Kannala,Marc Pollefeys,Baoquan Chen

Visual description, detection, and matching of keypoints in images are fundamental components of many computer vision problems, such as camera tracking and (re)localization. Recently, learning-based feature extractors on top of convolutional neural networks (CNNs) have achieved state-of-the-art performance. In this paper, we further explore the usage of CNN and present a new approach that ensembles randomly initialized CNNs without any training. Our observation is that the CNN architecture inherently extracts features with certain extents of robustness to viewpoint/illumination changes and thus, it can be regarded as visual descriptors. Consequently, randomized CNNs serve as descriptor extractors and a subsequent consensus mechanism detects keypoints using them. Such description and detection pipeline can be used to match keypoints in images and achieves higher generalization ability than the state-of-the-art methods in our experiments.

****

Visually-augmented pretrained language models for NLP Tasks without Images
Hangyu Guo,Kun Zhou,Xin Zhao,Qinyu Zhang,Ji-Rong Wen

Although pre-trained language models (PLMs) have shown impressive performance by text-only self-supervised training, they are found lack of visual semantics or commonsense, e.g., sizes, shapes and colors of commonplace objects. Existing solutions often rely on explicit images for visual knowledge augmentation (requiring time-consuming retrieval or generation), and they also conduct the augmentation for the whole input text, without considering whether it is actually needed in specific inputs or tasks. To address these issues, we propose a novel visually-augmented fine-tuning approach that can be generally applied to various PLMs or NLP tasks, without using any retrieved or generated images, namely VAWI. Specifically, we first identify the visually-hungry words (VH-words) from input text via a token selector, where three different methods have been proposed, including syntax-, attention- and learning-based strategies. Then, we adopts a fixed CLIP text encoder to generate the visually-augmented representations of these VH-words. As it has been pre-trained by visual-language alignment task on large-scale corpus, it is capable of injecting visual semantics into the aligned text representations. Finally, the visually-augmented features will be fused and transformed into several pre-designed visual prompts based on VH-words, which can be inserted into PLMs to enrich the visual semantics in word repersentations. We conduct extensive experiments on ten NLP tasks, i.e., GLUE benchmark, CommonsenseQA, CommonGen and SNLI-VE. Experimental results show that our approach can consistently improve the performance of BERT, RoBERTa, BART and T5 at different scales, and outperform several competitive baselines significantly. Besides, the generated visual prompts of our framework can also be used for parameter-efficient tuning, which boosts the performance of T5-3B. We will make our code, data, and models publicly available.

****

Improving Adversarial Robustness via Frequency Regularization

Binxiao Huang,Chaofan Tao,Rui Lin,Ngai Wong

Deep neural networks (DNNs) are incredibly vulnerable to crafted, human-imperceptible adversarial perturbations. While adversarial training (AT) has proven to be an effective defense approach, the properties of AT for robustness improvement remain an open issue. In this paper, we investigate AT from a spectral perspective, providing new insights into the design of effective defenses. Our analyses show that AT induces the deep model to focus more on the low-frequency region, which retains the shape-biased representations, to gain robustness. Further, we find that the spectrum of a white-box attack is primarily distributed in regions the model focuses on, and the perturbation attacks the spectral bands where the model is vulnerable. To train a model tolerant to frequency-varying perturbation, we propose a frequency regularization (FR) such that the spectral output inferred by an attacked input stays as close as possible to its natural input counterpart. Experiments demonstrate that FR and its weight averaging (WA) extension could significantly improve the robust accuracy by 1.14% ~ 4.57%, across multiple datasets (SVHN, CIFAR-10, CIFAR-100, and Tiny ImageNet), and various attacks (PGD, C&W, and Autoattack), without any extra data.

****

FastFill: Efficient Compatible Model Update

Florian Jaeckle,Fartash Faghri,Ali Farhadi,Oncel Tuzel,Hadi Pouransari

In many retrieval systems the original high dimensional data (e.g., images) is mapped to a lower dimensional feature through a learned embedding model. The task of retrieving the most similar data from a gallery set to a given query data is performed through similarity comparison on features. When the embedding model is updated, it might produce features that are not comparable/compatible with features already in the gallery computed with the old model. Subsequently, all features in the gallery need to be re-computed using the new embedding model -- a computationally expensive process called backfilling. Recently, compatible representation learning methods have been proposed to avoid back-filling. Despite their relative success, there is an inherent trade-off between new model performance and its compatibility with the old model. In this work, we introduce FastFill: a compatible model update process using feature alignment and policy based partial backfilling to promptly elevate retrieval performance. We show that previous backfilling strategies suffer from decreased performance and demonstrate the importance of both the training objective and the ordering in online partial backfilling. We propose a new training method for feature alignment between old and new embedding models using uncertainty estimation. Compared to previous works, we obtain significantly improved backfilling results on a variety of datasets: mAP on ImageNet (+4.4%), Places-365 (+2.7%), and VGG-Face2 (+1.3%). Further, we demonstrate that when updating a biased model with FastFill, the minority subgroup accuracy gap promptly vanishes with a small fraction of partial backfilling.

****

Learnable Graph Convolutional Attention Networks

Adrián Javaloy,Pablo Sanchez Martin,Amit Levi,Isabel Valera

Existing Graph Neural Networks (GNNs) compute the message exchange between nodes by either aggregating uniformly (convolving) the features of all the neighboring nodes, or by applying a non-uniform score (attending) to the features. Recent works have shown the strengths and weaknesses of the resulting GNN architectur es, respectively, GCNs and GATs. In this work, we aim at exploiting the strengths of both approaches to their full extent. To this end, we first introduce the graph convolutional attention layer (CAT), which relies on convolutions to compute the attention scores. Unfortunately, as in the case of GCNs and GATs, we show that there exists no clear winner between the three—neither theoretically nor in practice—as their performance directly depends on the nature of the data (i.e., of the graph and features). This result brings us to the main contribution of our work, the learnable graph convolutional attention network (L-CAT): a GNN architecture that automatically interpolates between GCN, GAT and CAT in each layer, by adding only two scalar parameters. Our results demonstrate that L-CAT is able to efficiently combine different GNN layers along the network, outperforming competing methods in a wide range of datasets, and resulting in a more robust model that reduces the need of cross-validating.

****

Indoor Localisation for Detecting Medication Use in Parkinson's Disease

Ferdian Jovan,Catherine Morgan,Ryan McConville,Emma Tonkin,Alan Whone,Ian Craddock

Parkinson’s disease (PD) is a slowly progressive debilitating neurodegenerative disease which is prominently characterised by motor symptoms. Indoor localisation, including its in-home mobility features, could provide a digital biomarker that can be used to quantify how mobility changes as this disease progresses. To improve the effectiveness of current methods for indoor localisation, a transformer-based approach utilising multiple modalities, Received Signal Strength Indicator (RSSI) and accelerometer data from wearable devices, which provide complementary views of movement, is proposed. To properly evaluate our proposed method, we use a free-living dataset where the movements and mobility are greatly varied and unstructured as expected in real-world conditions. 12 pairs of people (one with PD, and the other a control participant) lived for five days in a smart home with various sensors. Our evaluation on such a dataset, which includes subjects with and without PD, demonstrates that our proposed network outperforms the current state-of-the-art in indoor localisation. We also show how the accurate room-level localisation predictions can be transformed into in-home mobility features (i.e. room-to-room transition duration) which can be used to effectively classify whether the PD participant is taking their medications or withholding them (increasing their symptoms)

****

Scaffolding a Student to Instill Knowledge

Anil Kag,Durmus Alp Emre Acar,Aditya Gangrade,Venkatesh Saligrama

We propose a novel knowledge distillation (KD) method to selectively instill teacher knowledge into a student model motivated by situations where the student's capacity is significantly smaller than that of the teachers. In vanilla KD, the teacher primarily sets a predictive target for the student to follow, and we posit that this target is overly optimistic due to the student's lack of capacity. We develop a novel scaffolding scheme where the teacher, in addition to setting a predictive target, also scaffolds the student's prediction by censoring hard-to-learn examples. Scaffolding utilizes the same information as the teacher's soft-max predictions as inputs, and in this sense, our proposal can be viewed as a natural variant of vanilla KD. We show on synthetic examples that censoring hard-examples leads to smoothening the student's loss landscape so that the student encounters fewer local minima. As a result, it has good generalization properties. Against vanilla KD, we achieve improved performance and are comparable to more intrusive techniques that leverage feature matching on benchmark datasets.

****

User-Interactive Offline Reinforcement Learning

Phillip Swazinna,Steffen Udluft,Thomas Runkler

Offline reinforcement learning algorithms still lack trust in practice due to the risk that the learned policy performs worse than the original policy that generated the dataset or behaves in an unexpected way that is unfamiliar to the user. At the same time, offline RL algorithms are not able to tune their most import ant hyperparameter - the proximity of the learned policy to the original policy. We propose an algorithm that allows the user to tune this hyperparameter at run time, thereby addressing both of the above mentioned issues simultaneously. This allows users to start with the original behavior and grant successively greater deviation, as well as stopping at any time when the policy deteriorates or the behavior is too far from the familiar one.

****

No-regret Learning in Repeated First-Price Auctions with Budget Constraints

Rui Ai,Chang Wang,Chenchen Li,Jinshan Zhang,Wenhan Huang,Xiaotie Deng

Recently the online advertising market has exhibited a gradual shift from second-price auctions to first-price auctions. Although there has been a line of works concerning online bidding strategies in first-price auctions, it still remains open how to handle budget constraints in the problem. In the present paper, we initiate the study for a buyer with budgets to learn online bidding strategies in repeated first-price auctions. We propose an RL-based bidding algorithm against the optimal non-anticipating strategy under stationary competition. Our algorithm obtains
$\widetilde O(\sqrt T)$-regret if the bids are all revealed at the end of each round. With the restriction that the buyer only sees the winning bid after each round, our modified algorithm obtains $\widetilde O(T^{\frac{7}{12}})$-regret by techniques developed from survival analysis. Our analysis extends to the more general scenario where the buyer has any bounded instantaneous utility function with regrets of the same order.

****

Private and Efficient Meta-Learning with Low Rank and Sparse decomposition

Soumyabrata Pal,Prateek Varshney,Gagan Madan,Abhradeep Guha Thakurta,Gaurav Aggarwal,Pradeep Shenoy,Gaurav Srivastava,Prateek Jain

Meta-learning is critical for a variety of practical ML systems -- like personalized recommendations systems -- that are required to generalize to new tasks despite a small number of task-specific training points. Existing meta-learning techniques use two complementary approaches of either learning a low-dimensional representation of points for all tasks, or task-specific fine-tuning of a global model trained using all the tasks. In this work, we propose a novel meta-learning framework that combines both the techniques to enable handling of a large number of data-starved tasks. Our framework models network weights as a sum of low-rank and sparse matrices. This allows us to capture information from multiple domains together in the low-rank part while still allowing task specific personalization using the sparse part. We instantiate and study the framework in the linear setting, where the problem reduces to that of estimating the sum of a rank-r and a k-column sparse matrix using a small number of linear measurements. We propose an alternating minimization method with hard thresholding -- AMHT-LRS -- to learn the low-rank and sparse part effectively and efficiently. For the realizable, Gaussian data setting, we show that AMHT-LRS indeed solves the problem efficiently with nearly optimal samples. We extend AMHT-LRS to ensure that it preserves privacy of each individual user in the dataset, while still ensuring strong generalization with nearly optimal number of samples. Finally, on multiple datasets, we demonstrate that the framework allows personalized models to obtain superior performance in the data-scarce regime.

****

ωGNNs: Deep Graph Neural Networks Enhanced by Multiple Propagation Operators

Moshe Eliasof,Lars Ruthotto,Eran Treister

Graph Neural Networks (GNNs) are limited in their propagation operators. These operators often contain non-negative elements only and are shared across channels and layers, limiting the expressiveness of GNNs. Moreover, some GNNs suffer from over-smoothing, limiting their depth. On the other hand, Convolutional Neural Networks (CNNs) can learn diverse propagation filters, and phenomena like over-smoothing are typically not apparent in CNNs.

In this paper, we bridge this gap by incorporating trainable channel-wise weighting factors ω to learn and mix multiple smoothing and sharpening propagat ion operators at each layer. Our generic method is called ωGNN, and we study two variants: ωGCN and ωGAT.

For ωGCN, we theoretically analyse its behaviour and the impact of ω on the obtained node features. Our experiments confirm these findings, demonstrating and explaining how both variants do not over-smooth.

Additionally, we experiment with 15 real-world datasets on node- and graph-classification tasks, where our ωGCN and ωGAT perform better or on par with state-of-the-art methods.

****

Few-bit Backward: Quantized Gradients of Activation Functions for Memory Footprint Reduction

Georgii Sergeevich Novikov,Daniel Bershatsky,Julia Gusak,Alex Shonenkov,Denis Valerievich Dimitrov,Ivan Oseledets

Memory footprint is one of the main limiting factors for large neural network training. In backpropagation, one needs to store the input to each operation in the computational graph. Every modern neural network model has quite a few pointwise nonlinearities in its architecture, and such operation induces additional memory costs which --- as we show -- can be significantly reduced by quantization of the gradients.

We propose a systematic approach to compute optimal quantization of the retained gradients of the pointwise nonlinear functions with only a few bits per each element.

We show that such approximation can be achieved by computing optimal piecewise-constant approximation of the derivative of the activation function, which can be done by dynamic programming. The drop-in replacements are implemented for all popular nonlinearities and can be used in any existing pipeline. We confirm the memory reduction and the same convergence on several open benchmarks.

****

SLTUNET: A Simple Unified Model for Sign Language Translation

Biao Zhang,Mathias Müller,Rico Sennrich

Despite recent successes with neural models for sign language translation (SLT), translation quality still lags behind spoken languages because of the data scarcity and modality gap between sign video and text. To address both problems, we investigate strategies for cross-modality representation sharing for SLT. We propose SLTUNET, a simple unified neural model designed to support multiple SLTrelated tasks jointly, such as sign-to-gloss, gloss-to-text and sign-to-text translation. Jointly modeling different tasks endows SLTUNET with the capability to explore the cross-task relatedness that could help narrow the modality gap. In addition, this allows us to leverage the knowledge from external resources, such as abundant parallel data used for spoken-language machine translation (MT). We show in experiments that SLTUNET achieves competitive and even state-of-the-art performance on PHOENIX-2014T and CSL-Daily when augmented with MT data and equipped with a set of optimization techniques. We further use the DGS Corpus for end-to-end SLT for the first time. It covers broader domains with a significantly larger vocabulary, which is more challenging and which we consider to allow for a more realistic assessment of the current state of SLT than the former two. Still, SLTUNET obtains improved results on the DGS Corpus. Code is available at https://github.com/bzhangGo/sltunet.

****

Pruning by Active Attention Manipulation

Zahra Babaiee,Lucas Liebenwein,Ramin Hasani,Daniela Rus,Radu Grosu

Structured pruning of a CNN is typically achieved by applying discrete masks on the CNN's filter weights or activation maps, post-training. Here, we present a new filter-importance-scoring concept named pruning by active attention manipulation (PAAM), that sparsifies the CNN's set of filters through a particular attention mechanism, during-training. PAAM learns continuous filter scores from the filter weights by optimizing a cost function regularized by an additive term in the scores. As the filters are not independent, we use attention to dynamically learn their correlations. Moreover, by training the pruning scores of all layers simultaneously, PAAM can account for layer inter-dependencies, which is essential to finding a performant sparse sub-network. PAAM can also train and generate a pruned network from scratch in a straightforward, one-stage training process without requiring a pre-trained network. Finally, PAAM does not need layer-specific hyperparameters and pre-defined layer budgets, since it can implicitly determine the appropriate number of filters in each layer. Our experimental results on different network architectures suggest that PAAM outperforms state-of-the-art structured-pruning methods (SOTA). On CIFAR-10 dataset, without requiring a pre-trained baseline network, we obtain 1.02% and 1.19% accuracy gain and 52.3% and 54% parameters reduction, on ResNet56 and ResNet110, respectively. Similarly, on the ImageNet dataset, PAAM achieves 1.06% accuracy gain while pruning 51.1% of the parameters on ResNet50. For Cifar-10, this is better than the SOTA with a margin of 9.5% and 6.6%, respectively, and on ImageNet with a margin of 11%.

****

Robustness of Unsupervised Representation Learning without Labels

Aleksandar Petrov,Marta Kwiatkowska

Unsupervised representation learning leverages large unlabeled datasets and is competitive with supervised learning. But non-robust encoders may affect downstream task robustness. Recently, robust representation encoders have become of interest. Still, all prior work evaluates robustness using a downstream classification task. Instead, we propose a family of unsupervised robustness measures, which are model- and task-agnostic and label-free. We benchmark state-of-the-art representation encoders and show that none dominates the rest. We offer unsupervised extensions to the FGSM and PGD attacks. When used in adversarial training, they improve most unsupervised robustness measures, including certified robustness. We validate our results against a linear probe and show that, for MOCOv2, adversarial training results in 3 times higher certified accuracy, a 2-fold decrease in impersonation attack success rate and considerable improvements in certified robustness.

****

Understanding the Generalization of Adam in Learning Neural Networks with Proper Regularization

Difan Zou,Yuan Cao,Yuanzhi Li,Quanquan Gu

Adaptive gradient methods such as Adam have gained increasing popularity in deep learning optimization. However, it has been observed in many deep learning applications such as image classification, Adam can converge to a different solution with a worse test error compared to (stochastic) gradient descent, even with a fine-tuned regularization. In this paper, we provide a theoretical explanation for this phenomenon: we show that in the nonconvex setting of learning over-parameterized two-layer convolutional neural networks starting from the same random initialization, for a class of data distributions (inspired from image data), Adam and gradient descent (GD) can converge to different global solutions of the training objective with provably different generalization errors, even with weight decay regularization. In contrast, we show that if the training objective is convex, and the weight decay regularization is employed, any optimization algorithms including Adam and GD will converge to the same solution if the training is successful. This suggests that the generalization gap between Adam and SGD in the presence of weight decay regularization is closely tied to the nonconvex landscape of deep learning optimization, which cannot be covered by the recent neural tangent kernel (NTK) based analysis.

****

ACQL: An Adaptive Conservative Q-Learning Framework for Offline Reinforcement Learning

Kun Wu,Yinuo Zhao,Zhiyuan Xu,Zhengping Che,Chengxiang Yin,Chi Harold Liu,Qinru Qiu,Feifei Feng,Jian Tang

Offline Reinforcement Learning (RL), which relies only on static datasets without additional interactions with the environment, provides an appealing alternative to learning a safe and promising control policy. Most existing offline RL methods did not consider relative data quality and only crudely constrained the distribution gap between the learned policy and the behavior policy in general. Moreover, these algorithms cannot adaptively control the conservative level in more fine-grained ways, like for each state-action pair, leading to a performance drop, especially over highly diversified datasets. In this paper, we propose an Adaptive Conservative Q-Learning (ACQL) framework that enables more flexible control over the conservative level of Q-function for offline RL. Specifically, we present two adaptive weight functions to shape the Q-values for collected and out-of-distribution data. Then we discuss different conditions under which the conservative level of the learned Q-function changes and define the monotonicity with respect to data quality and similarity. Motivated by the theoretical analysis, we propose a novel algorithm with the ACQL framework, using neural networks as the adaptive weight functions. To learn proper adaptive weight functions, we design surrogate losses incorporating the conditions for adjusting conservative levels and a contrastive loss to maintain the monotonicity of adaptive weight functions. We evaluate ACQL on the commonly-used D4RL benchmark and conduct extensive ablation studies to illustrate the effectiveness and state-of-the-art performance compared to existing offline DRL baselines.

****

Fisher-Legendre (FishLeg) optimization of deep neural networks

Jezabel R Garcia,Federica Freddi,Stathi Fotiadis,Maolin Li,Sattar Vakili,Alberto Bernacchia,Guillaume Hennequin

Incorporating second-order gradient information (curvature) into optimization can dramatically reduce the number of iterations required to train machine learning models. In natural gradient descent, such information comes from the Fisher information matrix which yields a number of desirable properties. As exact natural gradient updates are intractable for large models, successful methods such as KFAC and sequels approximate the Fisher in a structured form that can easily be inverted. However, this requires model/layer-specific tensor algebra and certain approximations that are often difficult to justify. Here, we use ideas from Legendre-Fenchel duality to learn a direct and efficiently evaluated model for the product of the inverse Fisher with any vector, in an online manner, leading to natural gradient steps that get progressively more accurate over time despite noisy gradients. We prove that the resulting “Fisher-Legendre” (FishLeg) optimizer converges to a (global) minimum of non-convex functions satisfying the PL condition, which applies in particular to deep linear networks. On standard auto-encoder benchmarks, we show empirically that FishLeg outperforms standard first-order optimization methods, and performs on par with or better than other second-order methods, especially when using small batches. Thanks to its generality, we expect our approach to facilitate the handling of a variety neural network layers in future work.

****

A law of adversarial risk, interpolation, and label noise

Daniel Paleka,Amartya Sanyal

In supervised learning, it has been shown that label noise in the data can be interpolated without penalties on test accuracy. We show that interpolating label noise induces adversarial vulnerability, and prove the first theorem showing the relationship between label noise and adversarial risk for any data distribution. Our results are almost tight if we do not make any assumptions on the inductive bias of the learning algorithm. We then investigate how different components of this problem affect this result including properties of the distribution. We also discuss non-uniform label noise distributions; and prove a new theorem showing uniform label noise induces nearly as large an adversarial risk as the worst poisoning with the same noise rate. Then, we provide theoretical and empirical evidence that uniform label noise is more harmful than typical real-world label noise. Finally, we show how inductive biases amplify the effect of label noise and argue the need for future work in this direction.

****

Lossy Image Compression with Conditional Diffusion Models

Ruihan Yang,Stephan Mandt

Denoising diffusion models have recently marked a milestone in high-quality image generation. One may thus wonder if they are suitable for neural image compression. This paper outlines an end-to-end optimized image compression framework bas ed on a conditional diffusion model, drawing on the transform-coding paradigm. Besides the latent variables inherent to the diffusion process, this paper introduces an additional discrete "content" latent variable to condition the denoising process on. This variable is equipped with a hierarchical prior for entropy coding. The remaining "texture" latent variables characterizing the diffusion process are synthesized (either stochastically or deterministically) at decoding time. We furthermore show that the performance can be tuned toward perceptual metrics of interest. Our extensive experiments involving five datasets and 16 image perceptual quality assessment metrics show that our approach not only compares favorably in terms of rate and perceptual distortion tradeoffs but also shows robust performance under all metrics while other baselines show less consistent behavior.

****

ASIF: coupled data turns unimodal models to multimodal without training

Antonio Norelli,Marco Fumero,Valentino Maiorca,Luca Moschella,Emanuele Rodolà,Francesco Locatello

Aligning the visual and language spaces requires to train deep neural networks from scratch on giant multimodal datasets; CLIP trains both an image and a text encoder, while LiT manages to train just the latter by taking advantage of a pretrained vision network. In this paper, we show that sparse relative representations are sufficient to align text and images without training any network. Our method relies on readily available single-domain encoders (trained with or without supervision) and a modest (in comparison) number of image-text pairs. ASIF redefines what constitutes a multimodal model by explicitly disentangling memory from processing: here the model is defined by the embedded pairs of all the entries in the multimodal dataset, in addition to the parameters of the two encoders. Experiments on standard zero-shot visual benchmarks demonstrate the typical transfer ability of image-text models. Overall, our method represents a simple yet surprisingly strong baseline for foundation multi-modal models, raising important questions on their data efficiency and on the role of retrieval in machine learning.

****

Momentum Boosted Episodic Memory for Improving Learning in Long-Tailed RL Environments

Dolton Milagres Fernandes,Pramod Kaushik,Harsh Shukla,Bapi Raju Surampudi

Conventional Reinforcement Learning (RL) algorithms assume the distribution of the data to be uniform or mostly uniform. However, this is not the case with most real-world applications like autonomous driving or in nature, where animals roam. Some objects are encountered frequently, and most of the remaining experiences occur rarely; the resulting distribution is called \emph{Zipfian}. Taking inspiration from the theory of \emph{complementary learning systems}, an architecture for learning from Zipfian distributions is proposed where long tail states are discovered in an unsupervised manner and states along with their recurrent activation are kept longer in episodic memory. The recurrent activations are then reinstated from episodic memory using a similarity search, giving weighted importance. The proposed architecture yields improved performance in a Zipfian task over conventional architectures. Our method outperforms IMPALA by a significant margin of 20.3\% when maps/objects occur with a uniform distribution and by 50.2\% on the rarest 20\% of the distribution.

****

ProtoGNN: Prototype-Assisted Message Passing Framework for Non-Homophilous Graphs

Yanfei Dong,Mohammed Haroon Dupty,Lambert Deng,Yong Liang Goh,Wee Sun Lee

Many well-known Graph Neural Network (GNN) models assume the underlying graphs are homophilous, where nodes share similar features and labels with their neighbours. They rely on message passing that iteratively aggregates neighbour's features and often suffer performance degradation on non-homophilous graphs where useful information is hardly available in the local neighbourhood. In addition, earlier studies show that in some cases GNNs are even outperformed by Multi-Layer Perceptron, indicating insufficient exploitation of node feature information. Moti vated by the two limitations, we propose ProtoGNN, a novel message passing framework that augments existing GNNs by effectively combining node features with structural information. ProtoGNN learns multiple class prototypes for each class from raw node features with the slot-attention mechanism. These prototype representations are then transferred onto the structural node features with explicit message passing to all non-training nodes irrespective of distance. This form of message passing, from training nodes to class prototypes to non-training nodes, also serves as a shortcut that bypasses local graph neighbourhoods and captures global information. ProtoGNN is a generic framework which can be applied onto any of the existing GNN backbones to improve node representations when node features are strong and local graph information is scarce. We demonstrate through extensive experiments that ProtoGNN brings performance improvement to various GNN backbones and achieves state-of-the-art on several non-homophilous datasets.

****

MonoFlow: A Unified Generative Modeling Framework for GAN Variants
Mingxuan Yi,Zhanxing Zhu,Song Liu

Generative adversarial networks (GANs) play a minmax two-player game via adversarial training. The conventional understanding of adversarial training is that the discriminator is trained to estimate a divergence and the generator learns to minimize this divergence. We argue that despite the fact that many variants of GANs are developed following this paradigm, the existing theoretical understanding of GANs and the practical algorithms are inconsistent. In order to gain deeper theoretical insights and algorithmic inspiration for these GAN variants, we leverage Wasserstein gradient flows which characterize the evolution of particles in the sample space. Based on this, we introduce a unified generative modeling framework – MonoFlow: the particle evolution is rescaled via an arbitrary monotonically increasing mapping. Under our framework, adversarial training can be viewed as a procedure first obtaining MonoFlow's vector field via the discriminator and then the generator learns to parameterize the flow defined by the corresponding vector field. We also reveal the fundamental difference between variational divergence minimization and adversarial training. These analysis help us to identify what types of generator loss functions can lead to the successful training of GANs and suggest that GANs may have more loss designs beyond those developed in the literature, e.g., non-saturated loss, as long as they realize MonoFlow. Consistent empirical studies are also included to validate the effectiveness of our framework.

****

The Effective coalitions of Shapley value For Integrated Gradients
ShuYang Liu,Changjie Fan,Yu Xiong,Meng Wang,Yujing Hu,Tangjie Lv,Zixuan Chen,Runze Wu,Yang Gao

Many methods aim to explain deep neural networks (DNN) by attributing the prediction of DNN to its input features, like Integrated Gradients and Deep Shap, which both have critical baseline problems. Previous studies pursue a perfect but intractable baseline value, which is hard to find and has a very high computational cost, limiting the application range of these baseline methods. In this paper, we propose to find a set of baseline values corresponding to Shapley values which are easier to be found and have a lower computation cost. To solve computation dilemma of Shapley value, we propose Effective Shapley value (ES), a proportional sampling method to well simulate the ratios between the Shapley values of features and then propose Shapley Integrated Gradients (SIG) to combine Integrated Gradients with ES, to achieve a good balance between efficiency and effectiveness. Experiment results show that our ES method can well and stably approximate the ratios between Shapley values, and our SIG method has a much better and more accurate performance than common baseline values with similar computational costs.

****

Cold Rao-Blackwellized Straight-Through Gumbel-Softmax Gradient Estimator
Alexander Shekhovtsov

The problem of estimating the gradient of an expectation in discrete random variables arises in many applications: learning with discrete latent representations

, training neural networks with quantized weights, activations, conditional blocks, etc.
This work contributes to the development of the popular Gumbel-Softmax family of estimator, which is based on approximating argmax with a temperature-parametrized softmax. The state-of-the art in this family, the Gumbel-Rao estimator uses internal MC samples to reduce the variance.
We show that in the limit of zero temperature the internal integration has a closed form solution. The limit estimator, called ZGR, has a favorable bias and variance, is simple to implement and computationally inexpensive and is obviously free of the temperature hyperparameter. Furthermore, ZGR is unbiased for the class of quadratic functions of categorical variables and can be decomposed into a sum of two simple but not very well performing on their own estimators: the straight through estimator and the DARN estimator. Experiments thoroughly validate the method.
****
Tree-structure segmentation for logistic regression
Adrien Ehrhardt
The decision for a financial institution to accept or deny a loan is based on the probability of a client paying back their debt in time. This probability is given by a model such as a logistic regression, and estimated based on, e.g., the clients' characteristics, their credit history, the repayment performance. Historically, different models have been developed on different markets and/or credit products and/or addressed population. We show that this amounts to modelling default as a mixture model composed of a decision tree and logistic regression on its leaves (thereafter “logistic regression tree”). We seek to optimise this practice by considering the population to which a client belongs as a latent variable, which we will estimate. After exposing the context, the notations and the problem formalisation, we will conduct estimation using a Stochastic-Expectation-Maximisation (SEM) algorithm. We will finally show the performance on simulated data, and on real retail credit data from [COMPANY], as well as real open-source data.
****
Neural Image Compression with a Diffusion-based Decoder
Noor Fathima Khanum Mohamed Ghouse,Jens Petersen,Auke J. Wiggers,Tianlin Xu,Guillaume Sautiere
Diffusion probabilistic models have recently achieved remarkable success in generating high quality image and video data. In this work, we build on this class of generative models and introduce a method for lossy compression of high resolution images. The resulting codec, which we call \emph{DIffuson-based Residual Augmentation Codec (DIRAC)}, is the first neural codec to allow smooth traversal of the rate-distortion-perception tradeoff at test time, while obtaining competitive performance with GAN-based methods in perceptual quality. Furthermore, while sampling from diffusion probabilistic models is notoriously expensive, we show that in the compression setting the number of steps can be drastically reduced.
****
Learning ReLU networks to high uniform accuracy is intractable
Julius Berner,Philipp Grohs,Felix Voigtlaender
Statistical learning theory provides bounds on the necessary number of training samples needed to reach a prescribed accuracy in a learning problem formulated over a given target class. This accuracy is typically measured in terms of a generalization error, that is, an expected value of a given loss function. However, for several applications --- for example in a security-critical context or for problems in the computational sciences --- accuracy in this sense is not sufficient. In such cases, one would like to have guarantees for high accuracy on every input value, that is, with respect to the uniform norm. In this paper we precisely quantify the number of training samples needed for any conceivable training algorithm to guarantee a given uniform accuracy on any learning problem formulated over target classes containing (or consisting of) ReLU neural networks of a prescribed architecture. We prove that, under very general assumptions, the minimal number of training samples for this task scales exponentially both in the dept

h and the input dimension of the network architecture.

****

Caption supervision enables robust learners: a controlled study of distributionally robust model training

Benjamin Feuer,Ameya Joshi,Chinmay Hegde

Vision language models like CLIP are robust to natural distribution shifts, in part because CLIP learns on unstructured data using a technique called caption supervision; the model inteprets image-linked texts as ground-truth labels. In a carefully controlled comparison study, we show that CNNs trained on a standard cross-entropy loss can also benefit from caption supervision, in some cases even more than VL models, on the same data. To facilitate future experiments with high-accuracy caption-supervised models, we introduce CaptionNet, one piece of which is a class-balanced, fully supervised dataset with over 50,000 new human-labeled ImageNet-compliant samples which includes web-scraped captions. In a series of experiments on CaptionNet, we show how the choice of loss function, data filtration and supervision strategy enable robust computer vision.

****

Active Learning for Object Detection with Evidential Deep Learning and Hierarchical Uncertainty Aggregation

Younghyun Park,Wonjeong Choi,Soyeong Kim,Dong-Jun Han,Jaekyun Moon

Despite the huge success of object detection, the training process still requires an immense amount of labeled data. Although various active learning solutions for object detection have been proposed, most existing works do not take advantage of epistemic uncertainty, which is an important metric for capturing the usefulness of the sample. Also, previous works pay little attention to the attributes of each bounding box (e.g., nearest object, box size) when computing the informativeness of an image. In this paper, we propose a new active learning strategy for object detection that overcomes the shortcomings of prior works. To make use of epistemic uncertainty, we adopt evidential deep learning (EDL) and propose a new module termed model evidence head (MEH), that makes EDL highly compatible with object detection. Based on the computed epistemic uncertainty of each bounding box, we propose hierarchical uncertainty aggregation (HUA) for obtaining the informativeness of an image. HUA realigns all bounding boxes into multiple levels based on the attributes and aggregates uncertainties in a bottom-up order, to effectively capture the context within the image. Experimental results show that our method outperforms existing state-of-the-art methods by a considerable margin.

****

How Sharpness-Aware Minimization Minimizes Sharpness?

Kaiyue Wen,Tengyu Ma,Zhiyuan Li

Sharpness-Aware Minimization (SAM) is a highly effective regularization technique for improving the generalization of deep neural networks for various settings. However, the underlying working of SAM remains elusive because of various intriguing approximations in the theoretical characterizations. SAM intends to penalize a notion of sharpness of the model but implements a computationally efficient variant; moreover, a third notion of sharpness was used for proving generalization guarantees. The subtle differences in these notions of sharpness can indeed lead to significantly different empirical results. This paper rigorously nails down the exact sharpness notion that SAM regularizes and clarifies the underlying mechanism. We also show that the two steps of approximations in the original motivation of SAM individually lead to inaccurate local conclusions, but their combination accidentally reveals the correct effect, when full-batch gradients are applied. Furthermore, we also prove that the stochastic version of SAM in fact regularizes the third notion of sharpness mentioned above, which is most likely to be the preferred notion for practical performance. The key mechanism behind this intriguing phenomenon is the alignment between the gradient and the top eigenvector of Hessian when SAM is applied.

****

On discrete symmetries of robotics systems: A group-theoretic and data-driven analysis

Daniel Ordonez-Apraez,Mario Martin,Antonio Agudo,Francesc Moreno-Noguer

In this work, we study the Morphological Symmetries of dynamical systems with one or more planes of symmetry, a predominant feature in animal biology and robotic systems, characterized by the duplication and balanced distribution of body parts. These morphological symmetries imply that the system's dynamics are symmetric (or approximately symmetric), which in turn imprints symmetries in optimal control policies and in all proprioceptive and exteroceptive measurements related to the evolution of the system's dynamics. For data-driven methods, symmetry represents an inductive bias that justifies data augmentation and the construction of symmetric function approximators. To this end, we use Group Theory to present a theoretical and practical framework allowing for (1) the identification of the system's morphological symmetry Group $\G$, (2) the characterization of how the group acts upon the system state variables and any relevant measurement living in the Euclidean space, and (3) the exploitation of data symmetries through the use of $\G$-equivariant/$\G$-invariant Neural Networks, for which we present experimental results on synthetic and real-world applications, demonstrating how symmetry constraints lead to better sample efficiency and generalization while reducing the number of trainable parameters.

****

The Implicit Bias of Minima Stability in Multivariate Shallow ReLU Networks

Mor Shpigel Nacson,Rotem Mulayoff,Greg Ongie,Tomer Michaeli,Daniel Soudry

We study the type of solutions to which stochastic gradient descent converges when used to train a single hidden-layer multivariate ReLU network with the quadratic loss. Our results are based on a dynamical stability analysis. In the univariate case, it was shown that linearly stable minima correspond to network functions (predictors), whose second derivative has a bounded weighted L^1 norm. Notably, the bound gets smaller as the step size increases, implying that training with a large step size leads to `smoother' predictors. Here we generalize this result to the multivariate case, showing that a similar result applies to the Laplacian of the predictor. We demonstrate the tightness of our bound on the MNIST dataset, and show that it accurately captures the behavior of the solutions as a function of the step size. Additionally, we prove a depth separation result on the approximation power of ReLU networks corresponding to stable minima of the loss. Specifically, although shallow ReLU networks are universal approximators, we prove that stable shallow networks are not. Namely, there is a function that cannot be well-approximated by stable single hidden-layer ReLU networks trained with a non-vanishing step size. This is while the same function can be realized as a stable two hidden-layer ReLU network. Finally, we prove that if a function is sufficiently smooth (in a Sobolev sense) then it can be approximated arbitrarily well using shallow ReLU networks that correspond to stable solutions of gradient descent.

****

Consciousness-Aware Multi-Agent Reinforcement Learning

Jianzhun Shao,Hongchang Zhang,Yun Qu,Chang Liu,Shuncheng He,Yuhang Jiang,Xiangyang Ji

In cooperative multi-agent reinforcement learning, centralized training with decentralized execution (CTDE) shows great promise for a trade-off between independent Q-learning and joint action learning. However, vanilla CTDE methods assumed a fixed number of agents could hardly adapt to real-world scenarios where dynamic team compositions typically suffer from the dilemma of dramatic partial observability variance. Specifically, agents with extensive sight ranges are prone to be affected by trivial environmental substrates, dubbed the “attention distraction” issue; ones with limited observability can hardly sense their teammates, hindering the quality of cooperation. In this paper, we propose a Consciousness-Aware Multi-Agent reinforcement learning (CAMA) approach, which roots in a divide-and-conquer strategy to facilitate stable and sustainable teamwork. Concretely, CAMA targets dividing the input entities with controlled observability masks by an Entity Dividing Module (EDM) according to their execution relevance for consciousness learning. To tackle the attention distraction issue, the highly related entities are fed to a Consciousness Enhancement Module (CEM) for consciousness-a

ware representation extraction via action prediction with an inverse model. For better out-of-sight-range cooperation, the lowly related ones are compressed to brief messages by a Consciousness Replenishment Module (CRM) with a conditional mutual information estimator. Our CAMA outperforms the SOTA methods significantly on the challenging StarCraftII, MPE, and Traffic Junction benchmarks.
****

Better with Less: Data-Active Pre-training of Graph Neural Networks

Jiarong Xu,Renhong Huang,XIN JIANG,Yuxuan Cao,Carl Yang,Chunping Wang,Yang Yang

Recently, pre-training on graph neural networks (GNNs) has become an active research area and is used to learn transferable knowledge for downstream tasks with unlabeled data. The success of graph pre-training models is often attributed to the massive amount of input data. In this paper, however, we identify the curse of big data phenomenon in graph pre-training: more training samples and graph datasets do not necessarily lead to better performance. Motivated by this observation, we propose a better-with-less framework for graph pre-training: few, but carefully chosen data are fed into a GNN model to enhance pre-training. This novel pre-training pipeline is called the data-active graph pre-training (APT) framework, and is composed of a graph selector and a pre-training model. The graph selector chooses the most representative and instructive data points based on the inherent properties of graphs as well as the predictive uncertainty. The proposed uncertainty, as feedback from the pre-training model, measures the confidence level of the model to the data. When fed with the chosen data, on the other hand, the pre-training model grasps an initial understanding of the new, unseen data, and at the same time attempts to remember the knowledge learnt from the previous data. Therefore, the integration and interaction between these two components form a unified framework, in which graph pre-training is performed in a progressive way. Experiment results show that the proposed APT framework is able to obtain an efficient pre-training model with fewer training data and better downstream performance.
****

MAST: Masked Augmentation Subspace Training for Generalizable Self-Supervised Priors

Chen Huang,Hanlin Goh,Jiatao Gu,Joshua M. Susskind

Recent Self-Supervised Learning (SSL) methods are able to learn feature representations that are invariant to different data augmentations, which can then be transferred to downstream tasks of interest. However, different downstream tasks require different invariances for their best performance, so the optimal choice of augmentations for SSL depends on the target task. In this paper, we aim to learn self-supervised features that generalize well across a variety of downstream tasks (e.g., object classification, detection and instance segmentation) without knowing any task information beforehand. We do so by Masked Augmentation Subspace Training (or MAST) to encode in the single feature space the priors from different data augmentations in a factorized way. Specifically, we disentangle the feature space into separate subspaces, each induced by a learnable mask that selects relevant feature dimensions to model invariance to a specific augmentation. We show the success of MAST in jointly capturing generalizable priors from different augmentations, using both unique and shared features across the subspaces. We further show that MAST benefits from uncertainty modeling to reweight ambiguous samples from strong augmentations that may cause similarity mismatch in each subspace. Experiments demonstrate that MAST consistently improves generalization on various downstream tasks, while being task-agnostic and efficient during SSL. We also provide interesting insights about how different augmentations are related and how uncertainty reflects learning difficulty.
****

Graph-based Deterministic Policy Gradient for Repetitive Combinatorial Optimization Problems

Zhongyuan Zhao,Ananthram Swami,Santiago Segarra

We propose an actor-critic framework for graph-based machine learning pipelines with non-differentiable blocks, and apply it to repetitive combinatorial optimization problems (COPs) under hard constraints. Repetitive COP refers to problems to be solved repeatedly on graphs of the same or slowly changing topology but rapidly changing node or edge weights. Compared to one-shot COPs, repetitive COPs often rely on fast heuristics to solve one instance of the problem before the next one arrives, at the cost of a relatively large optimality gap. Through numerical experiments on several discrete optimization problems, we show that our approach can learn reusable node or edge representations to reduce the optimality gap of fast heuristics for independent repetitive COPs, and can optimize the long-term objectives for repetitive COPs embedded in graph-based Markov decision processes. Source code at https://github.com/XzrTGMu/twin-nphard

****

Lower Bounds on the Depth of Integral ReLU Neural Networks via Lattice Polytopes

Christian Alexander Haase,Christoph Hertrich,Georg Loho

We prove that the set of functions representable by ReLU neural networks with integer weights strictly increases with the network depth while allowing arbitrary width. More precisely, we show that $\lceil\log_2(n)\rceil$ hidden layers are indeed necessary to compute the maximum of n numbers, matching known upper bounds. Our results are based on the known duality between neural networks and Newton polytopes via tropical geometry. The integrality assumption implies that these Newton polytopes are lattice polytopes. Then, our depth lower bounds follow from a parity argument on the normalized volume of faces of such polytopes.

****

Contextual Transformer for Offline Reinforcement Learning

Runji Lin,Ye Li,Xidong Feng,Zhaowei Zhang,XIAN HONG WU FUNG,Haifeng Zhang,Jun Wang,Yali Du,Yaodong Yang

Recently, the pretrain-tuning paradigm in large-scale sequence models has made significant progress in Natural Language Processing and Computer Vision. However, such a paradigm is still hindered by intractable challenges in Reinforcement Learning (RL), including the lack of self-supervised large-scale pretraining methods based on offline data and efficient fine-tuning/prompt-tuning over unseen downstream tasks. In this work, we explore how prompts can help sequence-modeling-based offline Reinforcement Learning (offline-RL) algorithms. Firstly, we propose prompt tuning for offline RL, where a context vector sequence is concatenated with the input to guide the conditional generation. As such, we can pretrain a model on the offline dataset with supervised loss and learn a prompt to guide the policy to play the desired actions. Secondly, we extend the framework to the Meta-RL setting and propose Contextual Meta Transformer (CMT), which leverages the context among different tasks as the prompt to improve the performance on unseen tasks. We conduct extensive experiments across three different offline-RL settings: offline single-agent RL on the D4RL dataset, offline Meta-RL on the MuJoCo benchmark, and offline MARL on the SMAC benchmark. The results validate the strong performance, and generality of our methods.

****

Two-Dimensional Weisfeiler-Lehman Graph Neural Networks for Link Prediction

Yang Hu,Xiyuan Wang,Zhouchen Lin,Pan Li,Muhan Zhang

Link prediction is one important application of graph neural networks (GNNs). Most existing GNNs for link prediction are based on one-dimensional Weisfeiler-Lehman (1-WL) test. 1-WL-GNNs first compute node representations by iteratively passing neighboring node features to the center, and then obtain link representations by aggregating the pairwise node representations. As pointed out by previous works, this two-step procedure results in low discriminating power, as 1-WL-GNNs by nature learn node-level representations instead of link-level. In this paper, we study a completely different approach which can directly obtain node pair (link) representations based on \textit{two-dimensional Weisfeiler-Lehman (2-WL) tests}. 2-WL tests directly use links (2-tuples) as message passing units instead of nodes, and thus can directly obtain link representations. We theoretically analyze the expressive power of 2-WL tests to discriminate non-isomorphic links, and prove their superior link discriminating power than 1-WL. Based on different 2-WL variants, we propose a series of novel 2-WL-GNN models for link prediction. Experiments on a wide range of real-world datasets demonstrate their competitive performance to state-of-the-art baselines and superiority over plain 1-WL-GNNs.

****

Wasserstein Auto-encoded MDPs: Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees

Florent Delgrange,Ann Nowe,Guillermo Perez

Although deep reinforcement learning (DRL) has many success stories, the large-scale deployment of policies learned through these advanced techniques in safety-critical scenarios is hindered by their lack of formal guarantees. Variational Markov Decision Processes (VAE-MDPs) are discrete latent space models that provide a reliable framework for distilling formally verifiable controllers from any RL policy. While the related guarantees address relevant practical aspects such as the satisfaction of performance and safety properties, the VAE approach suffers from several learning flaws (posterior collapse, slow learning speed, poor dynamics estimates), primarily due to the absence of abstraction and representation guarantees to support latent optimization. We introduce the Wasserstein auto-encoded MDP (WAE-MDP), a latent space model that fixes those issues by minimizing a penalized form of the optimal transport between the behaviors of the agent executing the original policy and the distilled policy, for which the formal guarantees apply. Our approach yields bisimulation guarantees while learning the distilled policy, allowing concrete optimization of the abstraction and representation model quality. Our experiments show that, besides distilling policies up to 10 times faster, the latent model quality is indeed better in general. Moreover, we present experiments from a simple time-to-failure verification algorithm on the latent space. The fact that our approach enables such simple verification techniques highlights its applicability.

****

Towards graph-level anomaly detection via deep evolutionary mapping

Xiaoxiao Ma,Jian Yang,Jia Wu,Quan Z. Sheng

Graph-level anomaly detection aims at depicting anomalous individual graphs in a graph set. Due to its significance in various real-world application fields, such as identifying rare molecules in chemistry and detecting potential frauds in online social networks, graph-level anomaly detection has received great attention. In distinction from node- and edge-level anomaly detection that is devoted to identifying anomalies on a single graph, graph-level anomaly detection faces more significant challenges because both the intra- and inter-graph structural and attribute patterns need to be taken into account to distinguish anomalies that exhibit deviating structures, rare attributes or the both. Although deep graph representation learning shows effectiveness in fusing high-level representations and capturing characters of individual graphs, most of the existing works are defective in graph-level anomaly detection because of their limited capability in exploring information across graphs, the imbalanced data distribution of anomalies, and low interpretability of the black-box graph neural networks (GNNs). To bridge these gaps, we propose a novel deep evolutionary graph mapping framework named GmapAD, which can adaptively map each graph into a new feature space based on its similarity to a set of representative nodes chosen from the graph set. By automatically adjusting the candidate nodes using a specially designed evolutionary algorithm, anomalies and normal graphs are mapped to separate areas in the new feature space where a clear boundary between them can be learned. The selected candidate nodes can therefore be regarded as a benchmark for explaining anomalies because anomalies are more dissimilar/similar to the benchmark than normal graphs. Through our extensive experiments on nine real-world datasets, we demonstrate that exploring both intra- and inter-graph structural and attribute information are critical to spot anomalous graphs, and our framework outperforms the state of the art on all datasets used in the experiments.

****

Global Explainability of GNNs via Logic Combination of Learned Concepts

Steve Azzolin,Antonio Longa,Pietro Barbiero,Pietro Lio,Andrea Passerini

While instance-level explanation of GNN is a well-studied problem with plenty of approaches being developed, providing a global explanation for the behaviour of a GNN is much less explored, despite its potential in interpretability and debu gging. Existing solutions either simply list local explanations for a given class, or generate a synthetic prototypical graph with maximal score for a given class, completely missing any combinatorial aspect that the GNN could have learned. In this work, we propose GLGExplainer (Global Logic-based GNN Explainer), the first Global Explainer capable of generating explanations as arbitrary Boolean combinations of learned graphical concepts. GLGExplainer is a fully differentiable architecture that takes local explanations as inputs and combines them into a logic formula over graphical concepts, represented as clusters of local explanations.

Contrary to existing solutions, GLGExplainer provides accurate and human-interpretable global explanations that are perfectly aligned with ground-truth explanations (on synthetic data) or match existing domain knowledge (on real-world data). Extracted formulas are faithful to the model predictions, to the point of providing insights into some occasionally incorrect rules learned by the model, making GLGExplainer a promising diagnostic tool for learned GNNs.

****

BO-Muse: A Human expert and AI teaming framework for accelerated experimental design

Sunil Gupta,Alistair Shilton,Shannon Ryan,Arun Kumar Anjanapura Venkatesh,Majid Abdolshah,Hung Le,Santu Rana,Julian Berk,Mahad Rashid,Svetha Venkatesh

In this paper we introduce BO-Muse, a new approach to human-AI teaming for the optimisation of expensive blackbox functions. Inspired by the intrinsic difficulty of extracting expert knowledge and distilling it back into AI models and by observations of human behaviour in real-world experimental design, our algorithm lets the human expert take the lead in the experimental process. The human expert can use their domain expertise to its full potential, while the AI plays the role of a muse, injecting novelty and searching for areas of weakness to break the human out of over-exploitation induced by cognitive entrenchment. With mild assumptions, we show that our algorithm converges sub-linearly, at a rate faster than the AI or human alone. We validate our algorithm using synthetic data and with human experts performing real-world experiments.

****

Coordination Scheme Probing for Generalizable Multi-Agent Reinforcement Learning

Hao Ding,Chengxing Jia,Cong Guan,Feng Chen,Lei Yuan,Zongzhang Zhang,Yang Yu

Coordinating with previously unknown teammates without joint learning is a crucial need for real-world multi-agent applications, such as human-AI interaction. An active research topic on this problem is ad hoc teamwork, which improves agents' coordination ability in zero-shot settings. However, previous works can only solve the problem of a single agent's coordination with different teams, which is not in line with arbitrary group-to-group coordination in complex multi-agent scenarios. Moreover, they commonly suffer from limited adaptation ability within an episode in a zero-shot setting. To address these problems, we introduce the Coordination Scheme Probing (CSP) approach that applies a disentangled scheme probing module to represent and classify the newly arrived teammates beforehand with limited pre-collected episodic data and makes multi-agent control accordingly. To achieve generalization, CSP learns a meta-policy with multiple sub-policies that follow distinguished coordination schemes in an end-to-end fashion and automatically reuses it to coordinate with unseen teammates. Empirically, we show that the proposed method achieves remarkable performance compared to existing ad hoc teamwork and policy generalization methods in various multi-agent cooperative scenarios.

****

Generalization error bounds for Neural Networks with ReLU activation

Harsh Pandey,Amitabha Bagchi,Srikanta J. Bedathur,Arindam Bhattacharya

We show rigorous bounds on the generalization error for Neural Networks with ReLU activation under the condition that the network size doesn't grow with the training set size. In order to prove these bounds we weaken the notion of uniform stability of a learning algorithm in a probabilistic way by positing the notion of almost sure (a.s.) support stability and proving that if an algorithm has low enough a.s. support stability its generalization error tends to 0 as the trainin g set size increases. Further we show that for Stochastic Gradient Descent to be almost surely support stable we only need the loss function to be locally Lipschitz and locally smooth with probability 1, thereby showing low generalization error with weaker conditions than have been used in the literature. We then show that Neural Networks with ReLU activation and a doubly differentiable loss function possess these properties, thereby proving low generalization error. The caveat is that the size of NN must not grow with the size of the training set. Finally we present experimental evidence to validate our theoretical results.

****

Two Birds, One Stone: An Equivalent Transformation for Hyper-relational Knowledge Graph Modeling

Yu Liu,Shu Yang,Jingtao Ding,quanming yao,Yong Li

By representing knowledge in a primary triple associated with additional attribute value qualifiers, hyper-relational knowledge graph (HKG) that generalizes triple based knowledge graph (KG) has been attracting research attention recently. Compared with KG, HKG is enriched with the semantic difference between the primary triple and additional qualifiers as well as the structural connection between entities in hyper-relational graph structure. However, to model HKG, existing studies mainly focus on either semantic information or structural information therein, fail to capture both simultaneously. To tackle this issue, in this paper, we propose an equivalent transformation for HKG modeling, referred to as TransEQ. Specifically, the equivalent transformation transforms a HKG to a KG, which considers both semantic and structural characteristics. Then a generalized encoder-decoder framework is developed to bridge the modeling research between KG and HKG. In the encoder part, KG-based graph neural networks are leveraged for structural modeling; while in the decoder part, various HKG-based scoring functions are exploited for semantic modeling. Especially, we design the sharing embedding mechanism in the encoder-decoder framework with semantic relatedness captured. We further theoretically prove that TransEQ preserves complete information in the equivalent transformation, and also achieves full expressivity. Finally, extensive experiments on three benchmarks demonstrate the superior performance of TransEQ in terms of both effectiveness and efficiency. On the largest benchmark WikiPeople, TransEQ significantly improves the state-of-the-art models by 15% on MRR.

****

Gradient Gating for Deep Multi-Rate Learning on Graphs

T. Konstantin Rusch,Benjamin Paul Chamberlain,Michael W. Mahoney,Michael M. Bronstein,Siddhartha Mishra

We present Gradient Gating (G^2), a novel framework for improving the performance of Graph Neural Networks (GNNs). Our framework is based on gating the output of GNN layers with a mechanism for multi-rate flow of message passing information across nodes of the underlying graph. Local gradients are harnessed to further modulate message passing updates. Our framework flexibly allows one to use any basic GNN layer as a wrapper around which the multi-rate gradient gating mechanism is built. We rigorously prove that G^2 alleviates the oversmoothing problem and allows the design of deep GNNs. Empirical results are presented to demonstrate that the proposed framework achieves state-of-the-art performance on a variety of graph learning tasks, including on large-scale heterophilic graphs.

****

MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning

Mikayel Samvelyan,Akbir Khan,Michael D Dennis,Minqi Jiang,Jack Parker-Holder,Jakob Nicolaus Foerster,Roberta Raileanu,Tim Rocktäschel

Open-ended learning methods that automatically generate a curriculum of increasingly challenging tasks serve as a promising avenue toward generally capable reinforcement learning agents. Existing methods adapt curricula independently over either environment parameters (in single-agent settings) or co-player policies (in multi-agent settings). However, the strengths and weaknesses of co-players can manifest themselves differently depending on environmental features. It is thus crucial to consider the dependency between the environment and co-player when shaping a curriculum in multi-agent domains. In this work, we use this insight and extend Unsupervised Environment Design (UED) to multi-agent environments. We t hen introduce Multi-Agent Environment Design Strategist for Open-Ended Learning (MAESTRO), the first multi-agent UED approach for two-player zero-sum settings. MAESTRO efficiently produces adversarial, joint curricula over both environments and co-players and attains minimax-regret guarantees at Nash equilibrium. Our experiments show that MAESTRO outperforms a number of strong baselines on competitive two-player games, spanning discrete and continuous control settings.

****

Capturing the Motion of Every Joint: 3D Human Pose and Shape Estimation with Independent Tokens

Sen Yang,Wen Heng,Gang Liu,GUOZHONG LUO,Wankou Yang,Gang YU

In this paper we present a novel method to estimate 3D human pose and shape from monocular videos. This task requires directly recovering pixel-alignment 3D human pose and body shape from monocular images or videos, which is challenging due to its inherent ambiguity. To improve precision, existing methods highly rely on the initialized mean pose and shape as prior estimates and parameter regression with an iterative error feedback manner. In addition, video-based approaches model the overall change over the image-level features to temporally enhance the single-frame feature, but fail to capture the rotational motion at the joint level, and cannot guarantee local temporal consistency. To address these issues, we propose a novel Transformer-based model with a design of independent tokens. First, we introduce three types of tokens independent of the image feature: \textit{joint rotation tokens, shape token, and camera token}.

By progressively interacting with image features through Transformer layers, these tokens learn to encode the prior knowledge of human 3D joint rotations, body shape, and position information from large-scale data, and are updated to estimate SMPL parameters conditioned on a given image. Second, benefiting from the proposed token-based representation, we further use a temporal model to focus on capturing the rotational temporal information of each joint, which is empirically conducive to preventing large jitters in local parts. Despite being conceptually simple, the proposed method attains superior performances on the 3DPW and Human3.6M datasets. Using ResNet-50 and Transformer architectures, it obtains 42.0 mm error on the PA-MPJPE metric of the challenging 3DPW, outperforming state-of-the-art counterparts by a large margin. Code will be publicly available\footnote{\url{https://github.com/yangsenius/INT_HMR_Model}}.

****

Almost Linear Constant-Factor Sketching for ℓ_1 and Logistic Regression

Alexander Munteanu,Simon Omlor,David Woodruff

We improve upon previous oblivious sketching and turnstile streaming results for ℓ_1 and logistic regression, giving a much smaller sketching dimension achieving $O(1)$-approximation and yielding an efficient optimization problem in the sketch space. Namely, we achieve for any constant $c>0$ a sketching dimension of $\tilde{O}(d^{1+c})$ for ℓ_1 regression and $\tilde{O}(\mu d^{1+c})$ for logistic regression, where μ is a standard measure that captures the complexity of compressing the data. For ℓ_1-regression our sketching dimension is near-linear and improves previous work which either required $\Omega(\log d)$-approximation with this sketching dimension, or required a larger $\operatorname{poly}(d)$ number of rows. Similarly, for logistic regression previous work had worse $\operatorname{poly}(\mu d)$ factors in its sketching dimension. We also give a tradeoff that yields a $1+\varepsilon$ approximation in input sparsity time by increasing the total size to $(d\log(n)/\varepsilon)^{O(1/\varepsilon)}$ for ℓ_1 and to $(\mu d\log(n)/\varepsilon)^{O(1/\varepsilon)}$ for logistic regression. Finally, we show that our sketch can be extended to approximate a regularized version of logistic regression where the data-dependent regularizer corresponds to the variance of the individual logistic losses.

****

Neural-based classification rule learning for sequential data

Marine Collery,Philippe Bonnard,François Fages,Remy Kusters

Discovering interpretable patterns for classification of sequential data is of key importance for a variety of fields, ranging from genomics to fraud detection or more generally interpretable decision-making.

In this paper, we propose a novel differentiable fully interpretable method to discover both local and global patterns (i.e. catching a relative or absolute temporal dependency) for rule-based binary classification.
It consists of a convolutional binary neural network with an interpretable neural filter and a training strategy based on dynamically-enforced sparsity.
We demonstrate the validity and usefulness of the approach on synthetic datasets and on an open-source peptides dataset.
Key to this end-to-end differentiable method is that the expressive patterns used in the rules are learned alongside the rules themselves.
****
Q-learning Decision Transformer: Leveraging Dynamic Programming for Conditional Sequence Modelling in Offline RL
Taku Yamagata,Ahmed Khalil,Raul Santos-Rodriguez
Recent works have shown that tackling offline reinforcement learning (RL) with a conditional policy produces promising results. The Decision Transformer (DT) combines the conditional policy approach and a transformer architecture, showing competitive performance against several benchmarks. However, DT lacks stitching ability -- one of the critical abilities for offline RL to learn the optimal policy from sub-optimal trajectories. This issue becomes particularly significant when the offline dataset only contains sub-optimal trajectories.
On the other hand, the conventional RL approaches based on Dynamic Programming (such as Q-learning) do not have the same limitation; however, they suffer from unstable learning behaviours, especially when they rely on function approximation in an off-policy learning setting. In this paper, we propose the Q-learning Decision Transformer (QDT) to address the shortcomings of DT by leveraging the benefits of Dynamic Programming (Q-learning). It utilises the Dynamic Programming results to relabel the return-to-go in the training data to then train the DT with the relabelled data. Our approach efficiently exploits the benefits of these two approaches and compensates for each other's shortcomings to achieve better performance. We empirically show these in both simple toy environments and the more complex D4RL benchmark, showing competitive performance gains.

****
ϵ-Invariant Hierarchical Reinforcement Learning for Building Generalizable Policy
Yihan Li,Tianren Zhang,Jinsheng Ren,Feng Chen
Goal-conditioned Hierarchical Reinforcement Learning (HRL) has shown remarkable potential for solving complex control tasks. However, existing methods struggle in tasks that require generalization since the learned subgoals are highly task-specific and therefore hardly reusable. In this paper, we propose a novel HRL framework called \textit{ϵ-Invariant HRL} that uses abstract, task-agnostic subgoals reusable across tasks, resulting in a more generalizable policy. Although such subgoals are reusable, a transition mismatch problem caused by the inevitable incorrect value evaluation of subgoals can lead to non-stationary learning and even collapse. We mitigate this mismatch problem by training the high-level policy to be adaptable to the stochasticity manually injected into the low-level policy. As a result, our framework can leverage reusable subgoals to constitute a hierarchical policy that can effectively generalize to unseen new tasks. Theoretical analysis and experimental results in continuous control navigation tasks and challenging zero-shot generalization tasks show that our approach significantly outperforms state-of-the-art methods.
****
Learning Control by Iterative Inversion
Gal Leibovich,Guy Jacob,Or Avner,Gal Novik,Aviv Tamar
We formulate learning for control as an inverse problem - inverting a dynamical system to give the actions which yield desired behavior. The key challenge in this formulation is a distribution shift in the inputs to the function to be inverted - the learning agent can only observe the forward mapping (its actions' consequences) on trajectories that it can execute, yet must learn the inverse mapping for inputs-outputs that correspond to a different, desired behavior. We propos

e a general recipe for inverse problems with a distribution shift that we term $\textit{iterative inversion}$ - learn the inverse mapping under the current input distribution (policy), then use it on the desired output samples to obtain a new input distribution, and repeat.

As we show, iterative inversion can converge to the desired inverse mapping, but under rather strict conditions on the mapping itself.

We next apply iterative inversion to learn control. Our input is a set of demonstrations of desired behavior, given as video embeddings of trajectories (without actions), and our method iteratively learns to imitate trajectories generated by the current policy, perturbed by random exploration noise. We find that constantly adding the demonstrated trajectory embeddings as input to the policy when generating trajectories to imitate, a-la iterative inversion, we effectively steer the learning towards the desired trajectory distribution. To the best of our knowledge, this is the first exploration of learning control from the viewpoint of inverse problems, and the main advantage of our approach is simplicity - it does not require rewards, and only employs supervised learning, which can be easily scaled to use state-of-the-art trajectory embedding techniques and policy representations. Indeed, with a VQ-VAE embedding, and a transformer-based policy, we demonstrate non-trivial continuous control on several tasks. Further, we report an improved performance on imitating diverse behaviors compared to reward based methods.

****

DetectBench: An Object Detection Benchmark for OOD Generalization Algorithms

Fan Wu,Nanyang Ye,Lanqing HONG,Chensheng Peng,Bikang Pan,Huaihai Lyu,Heyuan Shi

The consensus about practical machine learning tasks, such as object detection, is still the test data are drawn from the same distribution as the training data, which is known as IID (Independent and Identically Distributed). However, it can not avoid being confronted with OOD (Out-of-Distribution) scenarios in real practice. It is risky to apply an object detection algorithm without figuring out its OOD generalization performance. On the other hand, a plethora of OOD generalization algorithms has been proposed to amortize the gap between the in-house and open-world performances of machine learning systems. However, their effectiveness was only demonstrated in the image classification tasks. It is still an opening question of how these algorithms perform on complex and practical tasks. In this paper, we first specify the setting of OOD-OD (OOD generalization object detection). Then, we propose DetectBench consisting of four OOD-OD benchmark datasets to evaluate various object detection and OOD generalization algorithms. From extensive experiments on DetectBench, we find that existing OOD generalization algorithms fail dramatically when applied to the more practical object detection tasks. This raises questions over the current progress on a large number of these algorithms and whether they can be effective in practice beyond simple toy examples. For future work, we sincerely hope that DetectBench can serve as a foothold for OOD-OD research.

****

Generalization Bounds with Arbitrary Complexity Measures

Paul Viallard,Rémi Emonet,Amaury Habrard,Emilie Morvant,Valentina Zantedeschi

In statistical learning theory, generalization bounds usually involve a complexity measure that is constrained by the considered theoretical framework. This limits the scope of such analysis, as in practical algorithms, other forms of regularization are used. Indeed, the empirical work of Jiang et al. (2019) shows that (I) common complexity measures (such as the VC-dimension) do not correlate with the generalization gap and that (ii) there exist arbitrary complexity measures that are better correlated with the generalization gap, but come without generalization guarantees. In this paper, we bridge the gap between this line of empirical works and generalization bounds of statistical learning theory. To do so, we leverage the framework of disintegrated PAC-Bayes bounds to derive a generalization bound that involves an arbitrary complexity measure. Our bound stands in probability jointly over the hypotheses and the learning sample, which allows us to improve the correlation between generalization gap and complexity, as the latter can be set to fit both the hypothesis class and the task.

****

Learning To Invert: Simple Adaptive Attacks for Gradient Inversion in Federated Learning

Ruihan Wu,Xiangyu Chen,Chuan Guo,Kilian Q Weinberger

Gradient inversion attack enables recovery of training samples from model updates in federated learning (FL) and constitutes a serious threat to data privacy. To mitigate this vulnerability, prior work proposed both principled defenses based on differential privacy, as well as heuristic defenses based on gradient compression as countermeasures. These defenses have so far been very effective, in particular those based on gradient compression that allow the model to maintain high accuracy while greatly reducing the attack's effectiveness. In this work, we argue that such findings do not accurately reflect the privacy risk in FL, and show that existing defenses can be broken by a simple adaptive attack that trains a model using auxiliary data to learn how to invert gradients on both vision and language tasks.

****

Leveraging Unlabeled Data to Track Memorization

Mahsa Forouzesh,Hanie Sedghi,Patrick Thiran

Deep neural networks may easily memorize noisy labels present in real-world data, which degrades their ability to generalize. It is therefore important to track and evaluate the robustness of models against noisy label memorization. We propose a metric, called $\textit{susceptibility}$, to gauge such memorization for neural networks. Susceptibility is simple and easy to compute during training. Moreover, it does not require access to ground-truth labels and it only uses unlabeled data. We empirically show the effectiveness of our metric in tracking memorization on various architectures and datasets and provide theoretical insights into the design of the susceptibility metric. Finally, we show through extensive experiments on datasets with synthetic and real-world label noise that one can utilize susceptibility and the overall training accuracy to distinguish models that maintain a low memorization on the training set and generalize well to unseen clean data.

****

CCIL: Context-conditioned imitation learning for urban driving

Ke Guo,Wei Jing,Wenxi Liu,Junbo Chen,Jia Pan

Imitation learning is a promising solution to the challenging autonomous urban driving task as experienced human drivers can effortlessly tackle highly complex driving scenarios. Behavior cloning is the most widely applied imitation learning approach in autonomous driving due to its exemption from potentially risky online interactions, but it suffers from the covariate shift issue. To mitigate this problem, we propose a context-conditioned imitation learning approach that learns a policy to map the context state into the ego vehicle's state instead of the typical formulation from both ego and context state to the ego action. Besides, to make full use of the spatial and temporal relations in the context to infer the ego future states, we design a novel policy network based on the Transformer, whose attention mechanism has demonstrated excellent performance in capturing relations. Finally, during evaluation, a linear quadratic controller is employed to produce smooth planning based on the predicted states from the policy network. Experiments on the real-world large-scale Lyft and nuPlan datasets demonstrate that our method can surpass the state-of-the-art method significantly.

****

Improving Continual Learning by Accurate Gradient Reconstructions of the Past

Erik Daxberger,Siddharth Swaroop,Kazuki Osawa,Rio Yokota,Richard E Turner,José Miguel Hernández-Lobato,Mohammad Emtiyaz Khan

Knowledge reuse is essential for continual learning, and current methods attempt to realize it through regularization or experience replay. These two strategies have complementary strengths, e.g., regularization methods are compact, but replay methods can mimic batch training more accurately. At present, little has been done to find principled ways to combine the two methods and current heuristics can give suboptimal performance. Here, we provide a principled approach to comb ine and improve them by using a recently proposed principle of adaptation, where the goal is to reconstruct the “gradients of the past”, i.e., to mimic batch training by estimating gradients from past data. Using this principle, we design a prior that provably gives better gradient reconstructions by utilizing two types of replay and a quadratic weight-regularizer. This improves performance on standard benchmarks such as Split CIFAR, Split TinyImageNet, and ImageNet-1000. Our work shows that a good combination of replay and regularizer-based methods can be very effective in reducing forgetting, and can sometimes even completely eliminate it.

****

Group-wise Verifiable Distributed Computing for Machine Learning under Adversarial Attacks

Sangwoo Hong,Heecheol Yang,Youngseok Yoon,Jungwoo Lee

Distributed computing has been a promising solution in machine learning to accelerate the training procedure on large-scale dataset by utilizing multiple workers in parallel. However, there remain two major issues that still need to be addressed: i) adversarial attacks from malicious workers, and ii) the effect of slow workers known as stragglers. In this paper, we tackle both problems simultaneously by proposing Group-wise Verifiable Coded Computing (GVCC), which leverages coding techniques and group-wise verification to provide robustness to adversarial attacks and resiliency to straggler effects in distributed computing. The key idea of GVCC is to verify a group of computation results from workers at a time, while providing resilience to stragglers through encoding tasks assigned to workers with Group-wise Verifiable Codes. Experimental results show that GVCC outperforms the existing methods in terms of overall processing time and verification time for executing matrix multiplication, which is a key computational component in machine learning and deep learning.

****

Extending graph transformers with quantum computed aggregation

Slimane Thabet,Romain Fouilland,Loic Henriet

Recently, efforts have been made in the community to design new Graph Neural Networks (GNN), as limitations of Message Passing Neural Networks became more apparent. This led to the appearance of Graph Transformers using global graph features such as Laplacian Eigenmaps. In our paper, we introduce a GNN architecture where the aggregation weights are computed using the long-range correlations of a quantum system. These correlations are generated by translating the graph topology into the interactions of a set of qubits in a quantum computer. The recent development of quantum processing units enables the computation of a new family of global graph features that would be otherwise out of reach for classical hardware. We give some theoretical insights about the potential benefits of this approach, and benchmark our algorithm on standard datasets. Although not being adapted to all datasets, our model performs similarly to standard GNN architectures, and paves a promising future for quantum enhanced GNNs.

****

Policy-Based Self-Competition for Planning Problems

Jonathan Pirnay,Quirin Göttl,Jakob Burger,Dominik Gerhard Grimm

AlphaZero-type algorithms may stop improving on single-player tasks in case the value network guiding the tree search is unable to approximate the outcome of an episode sufficiently well. One technique to address this problem is transforming the single-player task through self-competition. The main idea is to compute a scalar baseline from the agent’s historical performances and to reshape an episode’s reward into a binary output, indicating whether the baseline has been exceeded or not. However, this baseline only carries limited information for the agent about strategies how to improve. We leverage the idea of self-competition and directly incorporate a historical policy into the planning process instead of its scalar performance. Based on the recently introduced Gumbel AlphaZero (GAZ), we propose our algorithm GAZ ‘Play-to-Plan’ (GAZ PTP), in which the agent learns to find strong trajectories by planning against possible strategies of its past self. We show the effectiveness of our approach in two well-known combinatorial optimization problems, the Traveling Salesman Problem and the Job-Shop Scheduli ng Problem. With only half of the simulation budget for search, GAZ PTP consistently outperforms all selected single-player variants of GAZ.

****

Can Fair Federated Learning reduce the need for personalization?

Alex Iacob,Pedro Porto Buarque de Gusmao,Nicholas Donald Lane

Federated Learning (FL) allows edge devices to collaboratively train machine learning models without sharing local data. Since the data distribution varies across client partitions, the performance of the federated model on local data also varies. To solve this, fair FL approaches attempt to reduce the accuracy disparity between local partitions by emphasizing clients with larger losses during training; while local adaptation personalizes the federated model by re-training on local data to provide a device participation incentive in cases where a federated model underperforms relative to one trained locally---their accuracy difference is less than zero. This paper evaluates Q-Fair Federated Learning (Q-FFL) in this relative domain and determines whether it provides a better starting point for personalization or supplants it. Contrary to expectation, Q-FFL does not significantly reduce the number of underperforming clients in a language task while doubling them in an image recognition task. Furthermore, fairness levels which maintain average accuracy provide no benefit to relative accuracy in federated or adapted models. We postulate that Q-FFL is unsuitable for our goal since clients with highly accurate local models require the federated model to have a disproportionate local partition accuracy to receive a benefit. Instead, we propose using knowledge distillation during FL training to create models with a higher local accuracy floor without forfeiting the ceiling. Our preliminary evaluation shows a 50% reduction in underperforming clients in the language task with no increase in underperforming clients for the image task. Thus, we argue that this simple change represents a more promising avenue for reducing the need for personalization than fairness.

****

Out-of-Distribution Detection based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy

Jinsong Zhang,Qiang Fu,Xu Chen,Lun Du,Zelin Li,Gang Wang,xiaoguang Liu,Shi Han,Dongmei Zhang

Out-of-Distribution (OOD) detection is essential for safety-critical applications of deep neural networks. OOD detection is challenging since DNN models may produce very high logits value even for OOD samples. Hence, it is of great difficulty to discriminate OOD data by directly adopting Softmax on output logits as the confidence score. Differently, we detect the OOD sample with Hopfield energy in a store-then-compare paradigm. In more detail, penultimate layer outputs on the training set are considered as the representations of in-distribution (ID) data. Thus they can be transformed into stored patterns that serve as anchors to measure the discrepancy of unseen data for OOD detection. Starting from the energy function defined in Modern Hopfield Network for the discrepancy score calculation, we derive a simplified version SHE with theoretical analysis. In SHE, we utilize only one stored pattern to present each class, and these patterns can be obtained by simply averaging the penultimate layer outputs of training samples within this class. SHE has the advantages of hyperparameterfree
and high computational efficiency. The evaluations of nine widely-used OOD datasets show the promising performance of such a simple yet effective approach and its superiority over State-of-the-Art models. Code is available at https://github.com/zjs975584714/SHE ood detection.

****

Scaling Pareto-Efficient Decision Making via Offline Multi-Objective RL

Baiting Zhu,Meihua Dang,Aditya Grover

The goal of multi-objective reinforcement learning (MORL) is to learn policies that simultaneously optimize multiple competing objectives. In practice, an agent's preferences over the objectives may not be known apriori, and hence, we require policies that can generalize to arbitrary preferences at test time. In this work, we propose a new data-driven setup for offline MORL, where we wish to learn a preference-agnostic policy agent using only a finite dataset of offline demon strations of other agents and their preferences. The key contributions of this work are two-fold. First, we introduce D4MORL, (D)atasets for MORL that are specifically designed for offline settings. It contains 1.8 million annotated demonstrations obtained by rolling out reference policies that optimize for randomly sampled preferences on 6 MuJoCo environments with 2-3 objectives each. Second, we propose Pareto-Efficient Decision Agents (PEDA), a family of offline MORL algorithms that builds and extends Decision Transformers via a novel preference-and-return-conditioned policy. Empirically, we show that PEDA closely approximates the behavioral policy on the D4MORL benchmark and provides an excellent approximation of the Pareto-front with appropriate conditioning, as measured by the hypervolume and sparsity metrics.

****

Learning from Asymmetrically-corrupted Data in Regression for Sensor Magnitude

Takayuki Katsuki,Takayuki Osogami

This paper addresses a regression problem in which output label values represent the results of sensing the magnitude of a phenomenon. A low value of such labels can either mean that the actual magnitude of the phenomenon has been low or that the sensor has made an incomplete observation. This leads to a bias toward lower values in labels and its resultant learning because labels for incomplete observations are recorded as lower than those for typical observations, even if both have monitored similar phenomena. Moreover, because an incomplete observation does not provide any tags indicating incompleteness, we cannot eliminate or impute them. To address this issue, we propose a learning algorithm that explicitly models the incomplete observations to be corrupted with an asymmetric noise that always has a negative value. We show that our algorithm is unbiased with a regression learned from the uncorrupted data that does not involve incomplete observations. We demonstrate the advantages of our algorithm through numerical experiments.

****

NAGphormer: A Tokenized Graph Transformer for Node Classification in Large Graphs

Jinsong Chen,Kaiyuan Gao,Gaichao Li,Kun He

The graph Transformer emerges as a new architecture and has shown superior performance on various graph mining tasks. In this work, we observe that existing graph Transformers treat nodes as independent tokens and construct a single long sequence composed of all node tokens so as to train the Transformer model, causing it hard to scale to large graphs due to the quadratic complexity on the number of nodes for the self-attention computation. To this end, we propose a Neighborhood Aggregation Graph Transformer (NAGphormer) that treats each node as a sequence containing a series of tokens constructed by our proposed Hop2Token module. For each node, Hop2Token aggregates the neighborhood features from different hops into different representations and thereby produces a sequence of token vectors as one input. In this way, NAGphormer could be trained in a mini-batch manner and thus could scale to large graphs. Moreover, we mathematically show that as compared to a category of advanced Graph Neural Networks (GNNs), the decoupled Graph Convolutional Network, NAGphormer could learn more informative node representations from the multi-hop neighborhoods. Extensive experiments on benchmark datasets from small to large are conducted to demonstrate that NAGphormer consistently outperforms existing graph Transformers and mainstream GNNs. Code is available at https://github.com/JHL-HUST/NAGphormer.

****

Bayesian Oracle for bounding information gain in neural encoding models

Konstantin-Klemens Lurz,Mohammad Bashiri,Edgar Y. Walker,Fabian H. Sinz

In recent years, deep learning models have set new standards in predicting neural population responses. Most of these models currently focus on predicting the mean response of each neuron for a given input. However, neural variability around this mean is not just noise and plays a central role in several theories on neural computation. To capture this variability, we need models that predict full response distributions for a given stimulus. However, to measure the quality of such models, commonly used correlation-based metrics are not sufficient as they mainly care about the mean of the response distribution. An interpretable alternative evaluation metric for likelihood-based models is \textit{Information Gain} (IG) which evaluates the likelihood of a model relative to a lower and upper bound. However, while a lower bound is usually easy to obtain, constructing an upper bound turns out to be challenging for neural recordings with relatively low numbers of repeated trials, high (shared) variability, and sparse responses. In this work, we generalize the jack-knife oracle estimator for the mean---commonly used for correlation metrics---to a flexible Bayesian oracle estimator for IG based on posterior predictive distributions. We describe and address the challenges that arise when estimating the lower and upper bounds from small datasets. We then show that our upper bound estimate is data-efficient and robust even in the case of sparse responses and low signal-to-noise ratio. We further provide the derivation of the upper bound estimator for a variety of common distributions including the state-of-the-art zero-inflated mixture models, and relate IG to common mean-based metrics. Finally, we use our approach to evaluate such a mixture model resulting in 90% IG performance.

****

Near Optimal Private and Robust Linear Regression

Xiyang Liu,Prateek Jain,Weihao Kong,Sewoong Oh,Arun Suggala

We study the canonical statistical estimation problem of linear regression from n i.i.d. examples under (ε,δ)-differential privacy when a fraction of response variables are adversarially corrupted. We propose a variant of the popular differentially private stochastic gradient descent (DP-SGD) algorithm with two innovations: a full-batch gradient descent to improve sample complexity and a novel adaptive clipping to guarantee robustness. When there is no adversarial corruption, this algorithm improves upon the existing state-of-the-art approach and achieves near optimal sample complexity. Under label-corruption, this is the first efficient linear regression algorithm to provably guarantee both (ϵ,δ)-DP and robustness. Synthetic experiments confirm the superiority of our approach.

****

Inverse Learning with Extremely Sparse Feedback for Recommendation

Guanyu Lin,Yu Zheng,Chen Gao,Jianxin Chang,Yanan Niu,Yang Song,Zhiheng Li,Depeng Jin,Yong Li

Negative sampling is widely used in modern recommender systems, where negative data is randomly sampled from the whole item pool. However, such a strategy often introduces false-positive noises. Existing approaches about de-noising recommendation mainly focus on positive instances while ignoring the noise in the large amount of sampled negative feedback. In this paper, we propose a meta learning method to annotate the unlabeled data from loss and gradient perspectives, which considers the noises on both positive and negative instances. Specifically, we first propose $\textit{inverse dual loss}$ (IDL) to boost the true label learning and prevent the false label learning, based on the loss of unlabeled data towards true and false labels during the training process. To achieve more robust sampling on hard instances, we further propose $\textit{inverse gradient}$ (IG) to explore the correct updating gradient and adjust the updating based on meta learning. We conduct extensive experiments on a benchmark and an industrially collected dataset where our proposed method can significantly improve AUC by 9.25% against state-of-the-art methods. Further analysis verifies the proposed inverse learning is model-agnostic and can well annotate the labels combined with different recommendation backbones. The source code along with the best hyper-parameter settings is available at this link: https://anonymous.4open.science/r/InverseLearning-4F4F.

****

Instance-Specific Augmentation: Capturing Local Invariances

Ning Miao,Tom Rainforth,Emile Mathieu,Yann Dubois,Yee Whye Teh,Adam Foster,Hyunjik Kim

We introduce InstaAug, a method for automatically learning input-specific augmentations from data. Previous data augmentation methods have generally assumed ind

ependence between the original input and the transformation applied to that input. This can be highly restrictive, as the invariances that the augmentations are based on are themselves often highly input dependent; e.g., we can change a leaf from green to yellow while maintaining its label, but not a lime. InstaAug instead allows for input dependency by introducing an invariance module that maps inputs to tailored transformation distributions. It can be simultaneously trained alongside the downstream model in a fully end-to-end manner, or separately learned for a pre-trained model. We empirically demonstrate that InstaAug learns meaningful input-dependent augmentations for a wide range of transformation classes, which in turn provides better performance on both supervised and self-supervised tasks.
****

Spectral Subgraph Localization

Judith Hermanns,Amit Boyarski,Petros Petsinis,Alex M. Bronstein,Davide Mottin,Panagiotis Karras

Several graph mining problems are based on some variant of the subgraph isomorphism problem: Given two graphs, G and Q, does G contain a subgraph isomorphic to Q? As this problem is NP-hard, many methods avoid addressing it explicitly. In this paper, we propose a method that solves the problem by localizing, i.e., finding the position of, Q in G, by means of an alignment among graph spectra. Finding a node correspondence from Q to G thereafter is relegated to a separate task, as an instance of the graph alignment problem. We demonstrate that our spectral approach outperforms a baseline based on the state-of-the-art method for graph alignment in terms of accuracy on real graphs and scales to hundreds of nodes as no other method does.
****

Dynamical Signatures of Learning in Recurrent Networks

Marius Schneider,Andreea Lazar

Recurrent neural networks (RNNs) are powerful computational tools that operate best near the edge of chaos, where small perturbations in neuronal firing are transmitted between neurons with minimal amplification or loss. In this article, we depart from the observation that both stimulus and noise can be seen as perturbations to the intrinsic dynamics of a recurrent network, however stimulus information must be reliably preserved, while noise must be discarded. First, we show that self-organizing recurrent networks (SORNs) that learn the spatio-temporal structure of their inputs, increase their recurrent memory by preferentially propagating the relevant stimulus-specific structured signals, while becoming more robust to random perturbation. We find that the computational advantages gained through self-supervised learning are accompanied by a shift from critical to ordered dynamics, and that this dynamical shift varies with the structure of the stimulus. Next, we show that SORNs with subcritical dynamics can outperform their random RNNs counterparts with critical dynamics, on a range of tasks, including a temporal MNIST and a sequential shape-rotation task. Interestingly, when a shape is rotated, both the invariant (shape) and the variant (motion direction) aspects of the stimulus sequence are improved through learning in the subcritical SORNs. We propose that the shift in criticality is a signature of specialization and we expect it to be found in all cases in which general-purpose recurrent networks acquire self-correcting properties by internalizing the statistical structure of their inputs.
****

Shifts 2.0: Extending The Dataset of Real Distributional Shifts

Andrey Malinin,andreas athanasopoulos,Muhamed Barakovic,Meritxell Bach Cuadra,Mark Gales,Cristina Granziera,Mara Graziani,Nikolay Kartashev,Konstantinos Kyriakopoulos,Po-Jui Lu,Nataliia Molchanova,Antonis Nikitakis,Vatsal Raina,Francesco La Rosa,Eli Sivena,Vasileios Tsarsitalidis,Efi Tsompopoulou,Elena Volf

Distributional shift, or the mismatch between training and deployment data, is a significant obstacle to the usage of machine learning in high-stakes industrial applications, such as autonomous driving and medicine. This creates a need to be able to assess how robustly ML models generalize as well as the quality of their uncertainty estimates. Standard ML datasets do not allow these properties to be assessed, as the training, validation and test data are often identically distributed. Recently, a range of dedicated benchmarks have appeared, featuring both distributionally matched and shifted data. The Shifts dataset stands out in terms of the diversity of tasks and data modalities it features. Unlike most benchmarks, which are dominated by 2D image data, Shifts contains tabular weather forecasting, machine translation, and vehicle motion prediction tasks. This enables models to be assessed on a diverse set of industrial-scale tasks and either universal or directly applicable task-specific conclusions to be reached. In this paper, we extend the Shifts Dataset with two datasets sourced from industrial, high-risk applications of high societal importance. Specifically, we consider the tasks of segmentation of white matter Multiple Sclerosis lesions in 3D magnetic resonance brain images and the estimation of power consumption in marine cargo vessels. Both tasks feature ubiquitous distributional shifts and strict safety requirements due to the high cost of errors. These new datasets will allow researchers to explore robust generalization and uncertainty estimation in new situations. This work provides a description of the dataset and baseline results for both tasks.

****

Λ-DARTS: Mitigating Performance Collapse by Harmonizing Operation Selection among Cells

Sajad Movahedi,Melika Adabinejad,Ayyoob Imani,Arezou Keshavarz,Mostafa Dehghani,Azadeh Shakery,Babak N Araabi

Differentiable neural architecture search (DARTS) is a popular method for neural architecture search (NAS), which performs cell-search and utilizes continuous relaxation to improve the search efficiency via gradient-based optimization. The main shortcoming of DARTS is performance collapse, where the discovered architecture suffers from a pattern of declining quality during search. Performance collapse has become an important topic of research, with many methods trying to solve the issue through either regularization or fundamental changes to DARTS.

However, the weight-sharing framework used for cell-search in DARTS and the convergence of architecture parameters has not been analyzed yet. In this paper, we provide a thorough and novel theoretical and empirical analysis on DARTS and its point of convergence.

We show that DARTS suffers from a specific structural flaw due to its weight-sharing framework that limits the convergence of DARTS to saturation points of the softmax function. This point of convergence gives an unfair advantage to layers closer to the output in choosing the optimal architecture, causing performance collapse. We then propose two new regularization terms that aim to prevent performance collapse by harmonizing operation selection via aligning gradients of layers.

Experimental results on six different search spaces and three different datasets show that our method (Λ-DARTS) does indeed prevent performance collapse, providing justification for our theoretical analysis and the proposed remedy.

****

Efficient Hyperparameter Optimization Through Tensor Completion

Aaman Rebello,Kriton Konstantinidis,Yao Lei Xu,Danilo Mandic

Hyperparameter optimization is a prerequisite for state-of-the-art performance in machine learning, with current strategies including Bayesian optimisation, Hyperband, and evolutionary methods. Whereas such methods have been shown to improve performance, none of these is designed to explicitly take advantage of the underlying data structure. To this end, we introduce a completely different approach for hyperaparameter optimization, based on low-rank tensor completion. This is achieved by first forming a multi-dimensional tensor which comprises performance scores for different combinations of hyperparameters. Based on the realistic underlying assumption that the so-formed tensor has a low rank structure, this then allows for reliable estimates of the unobserved validation scores of combinations of hyperparameters to be obtained through tensor completion, and from only a fraction of known elements. Through extensive experimentation on various datasets and learning models, the proposed method is shown to exhibit competitive or superior performance to the state-of-the-art hyperparameter optimization strate gies. Distinctive advantages of the proposed method include its ability to simultaneously handle any hyperparameter type (e.g., kind of optimizer, number of neurons, number of layer, etc.), its relative simplicity compared to the competing methods, as well as the ability to suggest multiple optimal combinations of hyperparameters.

****

Learning Vortex Dynamics for Fluid Inference and Prediction

Yitong Deng,Hong-Xing Yu,Jiajun Wu,Bo Zhu

We propose a novel differentiable vortex particle (DVP) method to infer and predict fluid dynamics from a single video. Lying at its core is a particle-based latent space to encapsulate the hidden, Lagrangian vortical evolution underpinning the observable, Eulerian flow phenomena. Our differentiable vortex particles are coupled with a learnable, vortex-to-velocity dynamics mapping to effectively capture the complex flow features in a physically-constrained, low-dimensional space. This representation facilitates the learning of a fluid simulator tailored to the input video that can deliver robust, long-term future predictions. The value of our method is twofold: first, our learned simulator enables the inference of hidden physics quantities (e.g., velocity field) purely from visual observation; secondly, it also supports future prediction, constructing the input video's sequel along with its future dynamics evolution. We compare our method with a range of existing methods on both synthetic and real-world videos, demonstrating improved reconstruction quality, visual plausibility, and physical integrity.

****

Discovering Generalizable Multi-agent Coordination Skills from Multi-task Offline Data

Fuxiang Zhang,Chengxing Jia,Yi-Chen Li,Lei Yuan,Yang Yu,Zongzhang Zhang

Cooperative multi-agent reinforcement learning (MARL) faces the challenge of adapting to multiple tasks with varying agents and targets. Previous multi-task MARL approaches require costly interactions to simultaneously learn or fine-tune policies in different tasks. However, the situation that an agent should generalize to multiple tasks with only offline data from limited tasks is more in line with the needs of real-world applications. Since offline multi-task data contains a variety of behaviors, an effective data-driven approach is to extract informative latent variables that can represent universal skills for realizing coordination across tasks. In this paper, we propose a novel Offline MARL algorithm to Discover coordInation Skills (ODIS) from multi-task data. ODIS first extracts task-invariant coordination skills from offline multi-task data and learns to delineate different agent behaviors with the discovered coordination skills. Then we train a coordination policy to choose optimal coordination skills with the centralized training and decentralized execution paradigm. We further demonstrate that the discovered coordination skills can assign effective coordinative behaviors, thus significantly enhancing generalization to unseen tasks. Empirical results in cooperative MARL benchmarks, including the StarCraft multi-agent challenge, show that ODIS obtains superior performance in a wide range of tasks only with offline data from limited sources.

****

On student-teacher deviations in distillation: does it pay to disobey?

Vaishnavh Nagarajan,Aditya Krishna Menon,Srinadh Bhojanapalli,Hossein Mobahi,Sanjiv Kumar

Knowledge distillation has been widely-used to improve the performance of a ``student'' network by hoping to mimic the soft probabilities of a  ``teacher'' network. Yet, for self-distillation to work, the student {\em must} deviate from the teacher in some manner \citep{stanton21does}. We conduct a variety of experiments across image and language classification datasets to more precisely understand the nature of student-teacher deviations and how they relate to accuracy gains. Our first key empirical observation is that in a majority of our settings, the student underfits points that the teacher finds hard. Next, we find that student-teacher deviations during the \textit{initial} phase training are \textit{not}

crucial to get the benefits of distillation --- simply switching to distillation in the middle of training can recover a significant fraction of distillation's accuracy gains.
We then provide two parallel theoretical perspectives of student-teacher deviations, one casting distillation as a regularizer in eigenspace, and another as a denoiser of gradients. In both these views, we argue how our empirically reported student-teacher deviations may emerge, and how they may relate to generalization. Importantly, our analysis bridges key gaps between existing theory and practice by focusing on gradient descent and avoiding label noise assumptions.
****
Merging Models Pre-Trained on Different Features with Consensus Graph
Tengfei Ma,Trong Nghia Hoang,Jie Chen
Learning global models effectively on private and decentralized datasets has become an increasingly important challenge of machine learning when applied in practice. Federated Learning (FL) has recently emerged as a solution paradigm to address this challenge. In particular, the FL clients agree to a common model parameterization in advance, which can then be updated collaboratively via synchronous aggregation of their local model updates. However, such strong requirement of modeling homogeneity and synchronicity across clients makes FL inapplicable to many practical learning scenarios that cannot afford such requirements. For example, in distributed sensing, a network of heterogeneous sensors sample from different data modalities of the same phenomenon. Each sensor thus requires its own specialized model. Local learning therefore needs to happen in isolation but inference still requires merging the local models for better performance.

To enable this, we investigate a feature fusion approach that extracts local feature representations from local models and incorporates them into a global representation to train a more holistic predictive model. We study two key aspects of this feature incorporation. First, we develop an alignment algorithm that draws accurate correspondence between feature components which are arbitrarily arranged across clients. Next, we propose learning a consensus graph that captures the high-order interactions between these feature components, which reveals how data with heterogeneous features can be stitched together coherently to train a better model. The proposed framework is demonstrated on four real-life data sets including monitoring and predicting power grids and traffic networks.
****
Unsupervised Performance Predictor for Architecture Search
Xiangning Xie,Yanan Sun,Yuqiao Liu
Performance predictors can directly predict the performance value of given neural architectures without training, thus broadly being studied to alleviate the prohibitive cost of Neural Architecture Search (NAS). However, existing performance predictors still require training a large number of architectures from scratch to get their performance labels as the training dataset, which is still computationally expensive. To solve this issue, we develop an unsupervised performance predictor called USPP, which can avoid costly dataset construction by using existing fully-trained architectures. Specifically, a progressive domain-invariant feature extraction method is proposed to assist in extracting domain-invariant features due to the great transferability challenge caused by the rich domain-specific features. Furthermore, a learnable representation (denoted as operation embedding) is designed to replace the fixed encoding of the operations to transfer more knowledge about operations to the target search space. In experiments, we train the predictor by the labeled architectures in NAS-Bench-101 and predict the architectures in the DARTS search space. Compared with other state-of-the-art NAS methods, the proposed USPP only costs 0.02 GPU days but finds the architecture with 97.86% on CIFAR-10 and 96.50% top-1 accuracy on ImageNet.
****
Efficient recurrent architectures through activity sparsity and sparse back-propagation through time
Anand Subramoney,Khaleelulla Khan Nazeer,Mark Schöne,Christian Mayr,David Kappel
Recurrent neural networks (RNNs) are well suited for solving sequence tasks in r

esource-constrained systems due to their expressivity and low computational requirements. However, there is still a need to bridge the gap between what RNNs are capable of in terms of efficiency and performance and real-world application requirements. The memory and computational requirements arising from propagating the activations of all the neurons at every time step to every connected neuron, together with the sequential dependence of activations, contribute to the inefficiency of training and using RNNs. We propose a solution inspired by biological neuron dynamics that makes the communication between RNN units sparse and discrete. This makes the backward pass with backpropagation through time (BPTT) computationally sparse and efficient as well. We base our model on the gated recurrent unit (GRU), extending it with units that emit discrete events for communication triggered by a threshold so that no information is communicated to other units in the absence of events. We show theoretically that the communication between units, and hence the computation required for both the forward and backward passes, scales with the number of events in the network. Our model achieves efficiency without compromising task performance, demonstrating competitive performance compared to state-of-the-art recurrent network models in real-world tasks, including language modeling. The dynamic activity sparsity mechanism also makes our model well suited for novel energy-efficient neuromorphic hardware. Code is available at https://github.com/KhaleelKhan/EvNN/.

****

Quality-Similar Diversity via Population Based Reinforcement Learning

Shuang Wu,Jian Yao,Haobo Fu,Ye Tian,Chao Qian,Yaodong Yang,QIANG FU,Yang Wei

Diversity is a growing research topic in Reinforcement Learning (RL). Previous research on diversity has mainly focused on promoting diversity to encourage exploration and thereby improve quality (the cumulative reward), maximizing diversity subject to quality constraints, or jointly maximizing quality and diversity, known as the quality-diversity problem. In this work, we present the quality-similar diversity problem that features diversity among policies of similar qualities. In contrast to task-agnostic diversity, we focus on task-specific diversity defined by a set of user-specified Behavior Descriptors (BDs). A BD is a scalar function of a trajectory (e.g., the fire action rate for an Atari game), which delivers the type of diversity the user prefers. To derive the gradient of the user-specified diversity with respect to a policy, which is not trivially available, we introduce a set of BD estimators and connect it with the classical policy gradient theorem. Based on the diversity gradient, we develop a population-based RL algorithm to adaptively and efficiently optimize the population diversity at multiple quality levels throughout training. Extensive results on MuJoCo and Atari demonstrate that our algorithm significantly outperforms previous methods in terms of generating user-specified diverse policies across different quality levels.

****

PREDICTION OF TOURISM FLOW WITH SPARSE DATA INCORPORATING TOURIST GEOLOCATIONS

Julian Lemmel,Zahra Babaiee,Marvin Kleinlehner,Ivan Majic,Philipp Neubauer,Johannes Scholz,Radu Grosu,Sophie Neubauer

Modern tourism in the 21st century is facing numerous challenges. One of these challenges is the rapidly growing number of tourists in space-limited regions such as historical city centers, museums, or geographical bottlenecks like narrow valleys. In this context, a proper and accurate prediction of tourism volume and tourism flow within a certain area is important and critical for visitor management tasks such as sustainable treatment of the environment and prevention of overcrowding. Static flow control methods like conventional low-level controllers or limiting access to overcrowded venues could not solve the problem yet. In this paper, we empirically evaluate the performance of state-of-the-art deep-learning methods such as RNNs, GNNs, and Transformers as well as the classic statistical ARIMA method. Granular limited data supplied by a tourism region is extended by exogenous data such as geolocation trajectories of individual tourists, weath er and holidays. In the field of visitor flow prediction with sparse data, we are thereby capable of increasing the accuracy of our predictions, incorporating modern input feature handling as well as mapping geolocation data on top of discrete POI data.

****

Modeling the Uncertainty with Maximum Discrepant Students for Semi-supervised 2D Pose Estimation

Jiaqi Wu,Junbiao Pang,Qingming Huang

Semi-supervised pose estimation is a practically challenging task for computer vision. Although numerous excellent semi-supervised classification methods have emerged, these methods typically use confidence to evaluate the quality of pseudo-labels, which is difficult to achieve in pose estimation tasks. For example, in pose estimation, confidence represents only the possibility that a position of the heatmap is a keypoint, not the quality of that prediction. In this paper, we propose a simple yet efficient framework to estimate the quality of pseudo-labels in semi-supervised pose estimation tasks from the perspective of modeling the uncertainty of the pseudo-labels. Concretely, under the dual mean-teacher framework, we construct the two maximum discrepant students (MDSs) to effectively push two teachers to generate different decision boundaries for the same sample. Moreover, we create multiple uncertainties to assess the quality of the pseudo-labels. Experimental results demonstrate that our method improves the performance of semi-supervised pose estimation on three datasets.

****

UTS: When Monotonic Value Factorisation Meets Non-monotonic and Stochastic Targets

Zeyang Liu,Lipeng Wan,Xue Sui,Xingyu Chen,Xuguang Lan

Extracting decentralised policies from joint action-values is an attractive way to exploit centralised learning. It is possible to apply monotonic value factorisation to guarantee consistency between the centralised and decentralised policies. However, the best strategy for training decentralised policies when the target joint action-values are non-monotonic and stochastic is still unclear. We propose a novel value factorisation method named uncertainty-based target shaping (UTS) to solve this problem. UTS employs networks that estimate the reward and the following state's embedding, where the large prediction error indicates that the target is stochastic. By replacing deterministic targets for the suboptimal with the best per-agent values, we enforce that all shaped targets become a subset of the space that can be represented by monotonic value factorisation. Empirical results show that UTS outperforms state-of-the-art baselines on multiple benchmarks, including matrix games, predator-prey, and challenging tasks in the StarCraft II micromanagement.

****

Meta-learning with Auto-generated Tasks for Predicting Human Behaviour in Normal Form Games

Shaofei Chen,Jiaxing Chen,Peng Li,William Yuan,Zhen Zhen Hu

In recent years, machine learning methods have been successfully applied to predict human behaviour in strategic settings. However, as available data of human behaviour is not always large enough and people's reasoning processes in different types of games are various, it is challenging to acquire a satisfied prediction model. In this paper, we employ a meta-learning method to improve learning performance in predicting human behaviour in normal form games. In particular, we first design a deep neural network that captures mixed human behaviour features to model and be learned to get a underlying behavioural predictor. Then, using a dataset of experimental human behaviour, we apply unsupervised learning to generate tasks and use meta-learning to improve the learning proficiency. Experimental results show that our proposed meta-learning method with the designed neural network and auto-generated tasks considerably increases the prediction accuracy and significantly exceeds the previous state-of-the-art.

****

FairGrad: Fairness Aware Gradient Descent

Gaurav Maheshwari,Michaël Perrot

We address the problem of group fairness in classification, where the objective is to learn models that do not unjustly discriminate against subgroups of the population. Most existing approaches are limited to simple binary tasks or involve difficult to implement training mechanisms. This reduces their practical applicability. In this paper, we propose FairGrad, a method to enforce fairness based on a reweighting scheme that iteratively learns group specific weights based on whether they are advantaged or not. FairGrad is easy to implement and can accommodate various standard fairness definitions. Furthermore, we show that it is competitive with standard baselines over various datasets including ones used in natural language processing and computer vision.

****

Better Teacher Better Student: Dynamic Prior Knowledge for Knowledge Distillation

Martin Zong,Zengyu Qiu,Xinzhu Ma,Kunlin Yang,Chunya Liu,Jun Hou,Shuai Yi,Wanli Ouyang

Knowledge distillation (KD) has shown very promising capabilities in transferring learning representations from large models (teachers) to small models (students). However, as the capacity gap between students and teachers becomes larger, existing KD methods fail to achieve better results. Our work shows that the 'prior knowledge' is vital to KD, especially when applying large teachers. Particularly, we propose the dynamic prior knowledge (DPK), which integrates part of teacher's features as the prior knowledge before the feature distillation. This means that our method also takes the teacher's feature as `input', not just `target'. Besides, we dynamically adjust the ratio of the prior knowledge during the training phase according to the feature gap, thus guiding the student in an appropriate difficulty. To evaluate the proposed method, we conduct extensive experiments on two image classification benchmarks (i.e. CIFAR100 and ImageNet) and an object detection benchmark (\i.e. MS COCO). The results demonstrate the superiority of our method in performance under varying settings. Besides, our DPK makes the performance of the student model positively correlated with that of the teacher model, which means that we can further boost the accuracy of students by applying larger teachers. More importantly, DPK provides a fast solution in teacher model selection for any given model. Our codes will be publicly available for reproducibility.

****

Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow

Xingchao Liu,Chengyue Gong,qiang liu

We present rectified flow, a simple approach to learning (neural) ordinary differential equation (ODE) models to transport between two empirically observed distributions π_0 and π_1, hence providing a unified solution to generative modeling and domain transfer, among various other tasks involving distribution transport. The idea of rectified flow is to learn the ODE to follow the straight paths connecting the points drawn from π_0 and π_1 as much as possible. This is achieved by solving a straightforward nonlinear least squares optimization problem, which can be easily scaled to large models without introducing extra parameters beyond standard supervised learning. The straight paths are the shortest paths between two points, and can be simulated exactly without time discretization and hence yield computationally efficient models. We show that, by learning a rectified flow from data, we effectively turn an arbitrary coupling of π_0 and π_1 to a new deterministic coupling with provably non-increasing convex transport costs. In addition, with a ``reflow" procedure that iteratively learns a new rectified flow from the data bootstrapped from the previous one, we obtain a sequence of flows with increasingly straight paths, which can be simulated accurately with coarse time discretization in the inference phase. In empirical studies, we show that rectified flow performs superbly on image generation, image-to-image translation, and domain adaptation. In particular, on image gen eration and translation, our method yields nearly straight flows that give high quality results even with \emph{a single Euler discretization step}. Code is available at \url{https://github.com/gnobitab/RectifiedFlow}.

****

Inequality phenomenon in l_{∞}-adversarial training, and its unrealized threats

Ranjie Duan,YueFeng Chen,Yao Zhu,Xiaojun Jia,Rong Zhang,Hui Xue'

The appearance of adversarial examples raises attention from both academia and industry. Along with the attack-defense arms race, adversarial training is the most effective against adversarial examples.

However, we find inequality phenomena occur during the l_{∞}-adversarial training, that few features dominate the prediction made by the adversarially trained model. We systematically evaluate such inequality phenomena by extensive experiments and find such phenomena become more obvious when performing adversarial training with increasing adversarial strength (evaluated by ϵ). We hypothesize such inequality phenomena make l_{∞}-adversarially trained model less reliable than the standard trained model when few ``important features" are influenced. To validate our hypothesis, we proposed two simple attacks that either perturb or replace important features with noise or occlusion. Experiments show that l_{∞}-adversarially trained model can be easily attacked when the few important features are influenced.

Our work shed light on the limitation of the practicality of l_{∞}-adversarial training.

****

Tensor-Based Sketching Method for the Low-Rank Approximation of Data Streams.

Cuiyu Liu,Xiao Chuanfu,Mingshuo Ding,Chao Yang

Low-rank approximation in data streams is a fundamental and significant task in computing science, machine learning and statistics. Multiple streaming algorithms have emerged over years and most of them are inspired by randomized algorithms, more specifically, sketching methods. However, many algorithms are not able to leverage information of data streams and consequently suffer from low accuracy. Existing data-driven methods improve accuracy but the training cost is expensive in practice. In this paper, from a subspace perspective, we propose a tensor-based sketching method for low-rank approximation of data streams. The proposed algorithm fully exploits the structure of data streams and obtains quasi-optimal sketching matrices by performing tensor decomposition on training data. A series of experiments are carried out and show that the proposed tensor-based method can be more accurate and much faster than the previous work.

****

CRISP: Curriculum based Sequential neural decoders for Polar code family

S Ashwin Hebbar,Viraj Vivek Nadkarni,Ashok Vardhan Makkuva,Suma Bhat,Sewoong Oh,Pramod Viswanath

Polar codes are widely used state-of-the-art codes for reliable communication that have recently been included in the 5^{th} generation wireless standards (5G). However, there still remains room for design of polar decoders that are both efficient and reliable in the short blocklength regime. Motivated by recent successes of data-driven channel decoders, we introduce a novel $\textbf{ C}$ur${\textbf{RI}}$culum based $\textbf{S}$equential neural decoder for $\textbf{P}$olar codes (CRISP).

We design a principled curriculum, guided by information-theoretic insights, to train CRISP and show that it outperforms the successive-cancellation (SC) decoder and attains near-optimal reliability performance on the $\text{Polar}(16,32)$ and $\text{Polar}(22,64)$ codes.

The choice of the proposed curriculum is critical in achieving the accuracy gains of CRISP, as we show by comparing against other curricula. More notably, CRISP can be readily extended to Polarization-Adjusted-Convolutional (PAC) codes, where existing SC decoders are significantly less reliable. To the best of our knowledge, CRISP constructs the first data-driven decoder for PAC codes and attains near-optimal performance on the $\text{PAC}(16,32)$ code.

****

A Mathematical Framework for Characterizing Dependency Structures of Multimodal Learning

Weida Wang,Tao Shi,Yaoyuan Liang,Xiangxiang Xu,Fei Ma,Shao-Lun Huang,Lizhong Zheng

Dependency structures between modalities have been utilized explicitly and implicitly in multimodal learning to enhance classification performance, particularly when the training samples are insufficient. Recent efforts have concentrated on developing suitable dependence structures and applying them in deep neural networks, but the interplay between the training sample size and various structures has not received enough attention. To address this issue, we propose a mathematical framework that can be utilized to characterize conditional dependency structures in analytic ways. It provides an explicit description of the sample size in learning various structures in a non-asymptotic regime. Additionally, it demonstrates how task complexity and a fitness evaluation of conditional dependence structures affect the results. Furthermore, we develop an autonomous updated coefficient algorithm auto-CODES based on the theoretical framework and conduct experiments on multimodal emotion recognition tasks using the MELD and IEMOCAP datasets. The experimental results validate our theory and show the effectiveness of the proposed algorithm.

****

Language Models are Realistic Tabular Data Generators

Vadim Borisov,Kathrin Sessler,Tobias Leemann,Martin Pawelczyk,Gjergji Kasneci

Tabular data is among the oldest and most ubiquitous forms of data. However, the generation of synthetic samples with the original data’s characteristics remains a significant challenge for tabular data. While many generative models from the computer vision domain, such as variational autoencoders or generative adversarial networks, have been adapted for tabular data generation, less research has been directed towards recent transformer-based large language models (LLMs), which are also generative in nature. To this end, we propose GReaT (Generation of Realistic Tabular data), which exploits an auto-regressive generative LLM to sample synthetic and yet highly realistic tabular data. Furthermore, GReaT can model tabular data distributions by conditioning on any subset of features; the remaining features are sampled without additional overhead. We demonstrate the effectiveness of the proposed approach in a series of experiments that quantify the validity and quality of the produced data samples from multiple angles. We find that GReaT maintains state-of-the-art performance across numerous real-world and synthetic data sets with heterogeneous feature types coming in various sizes.

****

Data augmentation alone can improve adversarial training

Lin Li,Michael W. Spratling

Adversarial training suffers from the issue of robust overfitting, which seriously impairs its generalization performance. Data augmentation, which is effective at preventing overfitting in standard training, has been observed by many previous works to be ineffective in mitigating overfitting in adversarial training. This work proves that, contrary to previous findings, data augmentation alone can significantly boost accuracy and robustness in adversarial training. We find that the hardness and the diversity of data augmentation are important factors in combating robust overfitting. In general, diversity can improve both accuracy and robustness, while hardness can boost robustness at the cost of accuracy within a certain limit and degrade them both over that limit. To mitigate robust overfitting, we first propose a new crop transformation Cropshift with improved diversity compared to the conventional one (Padcrop). We then propose a new data augmentation scheme, based on Cropshift, with much improved diversity and well-balanced hardness. Empirically, our augmentation method achieves the state-of-the-art accuracy and robustness for data augmentations in adversarial training. Furthermore, it matches, or even exceeds when combined with weight averaging, the performance of the best contemporary regularization methods for alleviating robust overfitting.

****

Learning Diffusion Bridges on Constrained Domains

Xingchao Liu,Lemeng Wu,Mao Ye,qiang liu

Diffusion models have achieved promising results on generative learning recently. However, because diffusion processes are most naturally applied on the unconstrained Euclidean space R^d, key challenges arise for developing diffusion based models for learning data on constrained and structured domains. We present a simple and unified framework to achieve this that can be easily adopted to various types of domains, including product spaces of any type (be it bounded/unbounded, continuous/discrete, categorical/ordinal, or their mix). In our model, the diffusion process is driven by a drift force that is a sum of two terms: one singular force designed by $Doob's~ h$-$transform$ that ensures all outcomes of the process to belong to the desirable domain, and one non-singular neural force field that is trained to make sure the outcome follows the data distribution statistically. Experiments show that our methods perform superbly on generating tabular data, images, semantic segments and 3D point clouds.

****

Revisiting Uncertainty Estimation for Node Classification: New Benchmark and Insights

Gleb Bazhenov,Denis Kuznedelev,Andrey Malinin,Artem Babenko,Liudmila Prokhorenkova

Uncertainty estimation is an important task that can be essential for high-risk applications of machine learning. This problem is especially challenging for node-level prediction in graph-structured data, as the samples (nodes) are interdependent. Recently, several studies addressed node-level uncertainty estimation. However, there is no established benchmark for evaluating these methods in a unified setup covering diverse distributional shift. In this paper, we address this problem and propose such a benchmark together with a technique for the controllable generation of data splits with various types of distributional shift. Importantly, besides the standard feature-based distributional shift, we also consider shifts specifically designed for graph-structured data. In summary, our benchmark consists of several graph datasets equipped with various distributional shift on which we evaluate the robustness of models and uncertainty estimation performance. This allows us to compare existing solutions in a unified setup. Moreover, we decompose the current state-of-the-art Dirichlet-based framework and perform an ablation study on its components. In our experiments, we demonstrate that when faced with complex yet realistic distributional shift, most models fail to maintain high classification performance and consistency of uncertainty estimates with prediction errors. However, ensembling techniques help to partially overcome significant drops in performance and achieve better results than distinct models. Among single-pass models, Natural Posterior Network with GNN encoder achieves the best performance.

****

CUTS: Neural Causal Discovery from Irregular Time-Series Data

Yuxiao Cheng,Runzhao Yang,Tingxiong Xiao,Zongren Li,Jinli Suo,Kunlun He,Qionghai Dai

Causal discovery from time-series data has been a central task in machine learning. Recently, Granger causality inference is gaining momentum due to its good explainability and high compatibility with emerging deep neural networks. However, most existing methods assume structured input data and degenerate greatly when encountering data with randomly missing entries or non-uniform sampling frequencies, which hampers their applications in real scenarios. To address this issue, here we present CUTS, a neural Granger causal discovery algorithm to jointly impute unobserved data points and build causal graphs, via plugging in two mutually boosting modules in an iterative framework: (i) Latent data prediction stage: designs a Delayed Supervision Graph Neural Network (DSGNN) to hallucinate and register unstructured data which might be of high dimension and with complex distribution; (ii) Causal graph fitting stage: builds a causal adjacency matrix with imputed data under sparse penalty. Experiments show that CUTS effectively infers causal graphs from irregular time-series data, with significantly superior performance to existing methods. Our approach constitutes a promising step towards applying causal discovery to real applications with non-ideal observations.

****

PAVI: Plate-Amortized Variational Inference

Louis Rouillard,Thomas Moreau,Demian Wassermann

Given observed data and a probabilistic generative model, Bayesian inference aims at obtaining the distribution of a model’s latent parameters that could have yielded the data. This task is challenging for large population studies where thousands of measurements are performed over a cohort of hundreds of subjects, resulting in a massive latent parameter space. This large cardinality renders off-the-shelf Variational Inference (VI) computationally impractical.

In this work, we design structured VI families that can efficiently tackle large population studies. Our main idea is to share the parameterization and learning across the different i.i.d. variables in a generative model --symbolized by the model’s plates. We name this concept plate amortization, and illustrate the powerful synergies it entitles, resulting in expressive, parsimoniously parameterized and orders of magnitude faster to train large scale hierarchical variational distributions.

We illustrate the practical utility of PAVI through a challenging Neuroimaging example featuring a million latent parameters, demonstrating a significant step towards scalable and expressive Variational Inference.

****

Near-optimal Coresets for Robust Clustering

Lingxiao Huang,Shaofeng H.-C. Jiang,Jianing Lou,Xuan Wu

We consider robust clustering problems in $\mathbb{R}^d$, specifically k-clustering problems (e.g., k-Median and k-Means) with m \emph{outliers}, where the cost for a given center set $C \subset \mathbb{R}^d$ aggregates the distances from C to all but the furthest m data points, instead of all points as in classical clustering. We focus on the ϵ-coreset for robust clustering, a small proxy of the dataset that preserves the clustering cost within ϵ-relative error for all center sets. Our main result is an ϵ-coreset of size $O(m + \mathrm{poly}(k \epsilon^{-1}))$ that can be constructed in near-linear time. This significantly improves previous results, which either suffers an exponential dependence on $(m + k)$ [Feldman and Schulman, SODA'12], or has a weaker bi-criteria guarantee [Huang et al., FOCS'18]. Furthermore, we show this dependence in m is nearly-optimal, and the fact that it is isolated from other factors may be crucial for dealing with large number of outliers. We construct our coresets by adapting to the outlier setting a recent framework [Braverman et al., FOCS'22] which was designed for capacity-constrained clustering, overcoming a new challenge that the participating terms in the cost, particularly the excluded m outlier points, are dependent on the center set C. We validate our coresets on various datasets, and we observe a superior size-accuracy tradeoff compared with popular baselines including uniform sampling and sensitivity sampling. We also achieve a significant speedup of existing approximation algorithms for robust clustering using our coresets.

****

CLUTR: Curriculum Learning via Unsupervised Task Representation Learning

Abdus Salam Azad,Izzeddin Gur,Aleksandra Faust,Pieter Abbeel,Ion Stoica

Reinforcement Learning (RL) algorithms are often known for sample inefficiency and difficult generalization. Recently, Unsupervised Environment Design (UED) emerged as a new paradigm for zero-shot generalization by simultaneously learning a task distribution and agent policies on the sampled tasks. This is a non-stationary process where the task distribution evolves along with agent policies; creating an instability over time. While past works demonstrated the potential of such approaches, sampling effectively from the task space remains an open challenge, bottlenecking these approaches. To this end, we introduce CLUTR: a novel curriculum learning algorithm that decouples task representation and curriculum learning into a two-stage optimization. It first trains a recurrent variational autoencoder on randomly generated tasks to learn a latent task manifold. Next, a teacher agent creates a curriculum by optimizing a minimax REGRET-based objective o n a set of latent tasks sampled from this manifold. By keeping the task manifold fixed, we show that CLUTR successfully overcomes the non-stationarity problem and improves stability. Our experimental results show CLUTR outperforms PAIRED, a principled and popular UED method, in terms of generalization and sample efficiency in the challenging CarRacing and navigation environments: showing an 18x improvement on the F1 CarRacing benchmark. CLUTR also performs comparably to the non-UED state-of-the-art for CarRacing, outperforming it in nine of the 20 tracks. CLUTR also achieves a 33% higher solved rate than PAIRED on a set of 18 out-of-distribution navigation tasks.

****

Test-time Adaptation for Segmentation via Image Synthesis

Mihir Prabhudesai,Anirudh Goyal,Sujoy Paul,Mehdi S. M. Sajjadi,Sjoerd van Steenkiste,Gaurav Aggarwal,Thomas Kipf,Deepak Pathak,Katerina Fragkiadaki

We consider the problem of segmenting scenes into constituent objects and their parts. Current supervised visual detectors, though impressive within their training distribution, often fail to segment out-of-distribution scenes into their constituent entities. Recent test-time adaptation methods use auxiliary self-supervised losses to adapt the network parameters to each test example independently and have shown promising results towards generalization outside the training distribution for the task of image classification. In our work, we find evidence that these losses can be insufficient for instance segmentation tasks, without also considering architectural inductive biases. For image segmentation, recent slot-centric generative models break such dependence on supervision by attempting to segment scenes into entities in a self-supervised manner by reconstructing pixels. Drawing upon these two lines of work, we propose Generating Fast and Slow Networks (GFS-Nets), a semi-supervised instance segmentation model equipped with a slot-centric image or point-cloud rendering component that is adapted per scene at test time through gradient descent on reconstruction or novel view synthesis objectives. We show that test-time adaptation greatly improves segmentation in out-of-distribution scenes. We evaluate GFS-Nets in several 3D and 2D scene segmentation benchmarks and show substantial out-of-distribution performance improvements against state-of-the-art supervised feed forward detectors and self-supervised domain adaptation models.

****

On the Importance of In-distribution Class Prior for Out-of-distribution Detection

Xue Jiang,Feng Liu,Zhen Fang,Hong Chen,Tongliang Liu,Feng Zheng,Bo Han

Given a pre-trained in-distribution (ID) model, the task of inference-time out-of-distribution (OOD) detection methods aims to recognize upcoming OOD data in inference time. However, some representative methods share an unproven assumption that the probability that OOD data belong to every ID class should be the same, i.e., probabilities that OOD data belong to ID classes form a uniform distribution. In this paper, we theoretically and empirically show that this assumption makes these methods incapable of recognizing OOD data when the ID model is trained with class-imbalanced data. Fortunately, by analyzing the causal relations between ID/OOD classes and features, we identify several common scenarios where probabilities that OOD data belong to ID classes should be the ID-class-prior distribution. Based on the above finding, we propose two effective strategies to modify previous inference-time OOD detection methods: 1) if they explicitly use the uniform distribution, we can replace the uniform distribution with the ID-class-prior distribution; 2) otherwise, we can reweight their scores according to the similarity between the ID-class-prior distribution and the softmax outputs of the pre-trained model. Extensive experiments show that both strategies significantly improve the accuracy of recognizing OOD data when the ID model is pre-trained with imbalanced data. As a highlight, when evaluating on the iNaturalist dataset, our method can achieve ~36% increase on AUROC and ~61% decrease on FPR95, compared with the original Energy method, reflecting the importance of ID-class prior in the OOD detection, which lights up a new road to study this problem.

****

Quantized Compressed Sensing with Score-Based Generative Models

Xiangming Meng,Yoshiyuki Kabashima

We consider the general problem of recovering a high-dimensional signal from noisy quantized measurements. Quantization, especially coarse quantization such as 1-bit sign measurements, leads to severe information loss and thus a good prior knowledge of the unknown signal is helpful for accurate recovery. Motivated by the power of score-based generative models (SGM, also known as diffusion models) in capturing the rich structure of natural signals beyond simple sparsity, we propose an unsupervised data-driven approach called quantized compressed sensing with SGM (QCS-SGM), where the prior distribution is modeled by a pre-trained SGM. To perform posterior sampling, an annealed pseudo-likelihood score called ${\textit{noise perturbed pseudo-likelihood score}}$ is introduced and combined with the prior score of SGM. The proposed QCS-SGM applies to an arbitrary number of quantization bits. Experiments on a variety of baseline datasets demonstrate that the proposed QCS-SGM significantly outperforms existing state-of-the-art algorithms by a large margin for both in-distribution and out-of-distribution samples. Moreover, as a posterior sampling method, QCS-SGM can be easily used to obtain confidence intervals or uncertainty estimates of the reconstructed results. $\textit{The code is available at}$ https://github.com/mengxiangming/QCS-SGM.

****

Unbiased Representation of Electronic Health Records for Patient Outcome Prediction

Lucas Jing Liu,Victor Ortiz,Javier A Neyra,Jin Chen

Fairness is one of the newly emerging focuses for building trustworthy artificial intelligence (AI) models. One of the reasons resulting in an unfair model is the algorithm bias towards different groups of samples. A biased model may benefit certain groups but disfavor others. As a result, leaving the fairness problem unresolved might have a significant negative impact, especially in the context of healthcare applications. Integrating both domain-specific and domain-invariant representations, we propose a masked triple attention transformer encoder (MTATE) to learn unbiased and fair data representations of different subpopulations. Specifically, MTATE includes multiple domain classifiers and uses three attention mechanisms to effectively learn the representations of diverse subpopulations.

In the experiment on real-world healthcare data, MTATE performed the best among the compared models regarding overall performance and fairness.

****

Valid P-Value for Deep Learning-driven Salient Region

Miwa Daiki,Vo Nguyen Le Duy,Ichiro Takeuchi

Various saliency map methods have been proposed to interpret and explain predictions of deep learning models. Saliency maps allow us to interpret which parts of the input signals have a strong influence on the prediction results. However, since a saliency map is obtained by complex computations in deep learning models, it is often difficult to know how reliable the saliency map itself is. In this study, we propose a method to quantify the reliability of a saliency region in the form of p-values. Our idea is to consider a saliency map as a selected hypothesis by the trained deep learning model and employ the selective inference framework. The proposed method provably provides a valid p-value for the detected salient region, i.e., we can provably control the false positive rate of the detected salient region. We demonstrate the validity of the proposed method through numerical examples in synthetic and real datasets. Furthermore, we develop a Keras-based framework for conducting the proposed selective inference for a wide class of CNNs without additional implementation cost.

****

Unsupervised Semantic Segmentation with Self-supervised Object-centric Representations

Andrii Zadaianchuk,Matthaeus Kleindessner,Yi Zhu,Francesco Locatello,Thomas Brox

In this paper, we show that recent advances in self-supervised representation learning enable unsupervised object discovery and semantic segmentation with a performance that matches the state of the field on supervised semantic segmentation 10 years ago. We propose a methodology based on unsupervised saliency masks and self-supervised feature clustering to kickstart object discovery followed by training a semantic segmentation network on pseudo-labels to bootstrap the system on images with multiple objects. We show that while being conceptually simple our proposed baseline is surprisingly strong. We present results on PASCAL VOC that go far beyond the current state of the art (50.0 mIoU), and we report for the first time results on MS COCO for the whole set of 81 classes: our method discovers 34 categories with more than 20% IoU, while obtaining an average IoU of 19.6 for all 81 categories.

****

Pre-training Protein Structure Encoder via Siamese Diffusion Trajectory Prediction

Zuobai Zhang,Minghao Xu,Aurelie Lozano,Vijil Chenthamarakshan,Payel Das,Jian Tang

Due to the determining role of protein structures on diverse protein functions, pre-training representations of proteins on massive unlabeled protein structures has attracted rising research interests. Among recent efforts on this direction, mutual information (MI) maximization based methods have gained the superiority on various downstream benchmark tasks. The core of these methods is to design correlated views that share common information about a protein. Previous view designs focus on capturing structural motif co-occurrence on the same protein structure, while they cannot capture detailed atom/residue interactions. To address this limitation, we propose the Siamese Diffusion Trajectory Prediction (SiamDiff) method. SiamDiff builds a view as the trajectory that gradually approaches protein native structure from scratch, which facilitates the modeling of atom/residue interactions underlying the protein structural dynamics. Specifically, we employ the multimodal diffusion process as a faithful simulation of the structure-sequence co-diffusion trajectory, where rich patterns of protein structural changes are embedded. On such basis, we design a principled theoretical framework to maximize the MI between correlated multimodal diffusion trajectories. We study the effectiveness of SiamDiff on both residue-level and atom-level structures. On the EC and ATOM3D benchmarks, we extensively compare our method with previous protein structure pre-training approaches. The experimental results verify the consistently superior or competitive performance of SiamDiff on all benchmark tasks compared to existing baselines. The source code will be made public upon acceptance.

****

Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning

Hao He,Kaiwen Zha,Dina Katabi

Indiscriminate data poisoning attacks are quite effective against supervised learning. However, not much is known about their impact on unsupervised contrastive learning (CL). This paper is the first to consider indiscriminate poisoning attacks of contrastive learning. We propose Contrastive Poisoning (CP), the first effective such attack on CL. We empirically show that Contrastive Poisoning, not only drastically reduces the performance of CL algorithms, but also attacks supervised learning models, making it the most generalizable indiscriminate poisoning attack. We also show that CL algorithms with a momentum encoder are more robust to indiscriminate poisoning, and propose a new countermeasure based on matrix completion. Code is available at: https://github.com/kaiwenzha/contrastive-poisoning.

****

Decompositional Generation Process for Instance-Dependent Partial Label Learning

Congyu Qiao,Ning Xu,Xin Geng

Partial label learning (PLL) is a typical weakly supervised learning problem, where each training example is associated with a set of candidate labels among which only one is true. Most existing PLL approaches assume that the incorrect labels in each training example are randomly picked as the candidate labels and model the generation process of the candidate labels in a simple way. However, these approaches usually do not perform as well as expected due to the fact that the generation process of the candidate labels is always instance-dependent. Therefore, it deserves to be modeled in a refined way. In this paper, we consider ins tance-dependent PLL and assume that the generation process of the candidate labels could decompose into two sequential parts, where the correct label emerges first in the mind of the annotator but then the incorrect labels related to the feature are also selected with the correct label as candidate labels due to uncertainty of labeling. Motivated by this consideration, we propose a novel PLL method that performs Maximum A Posterior(MAP) based on an explicitly modeled generation process of candidate labels via decomposed probability distribution models. Extensive experiments on manually corrupted benchmark datasets and real-world datasets validate the effectiveness of the proposed method.

****

Multimodal Masked Autoencoders Learn Transferable Representations

Xinyang Geng,Hao Liu,Lisa Lee,Dale Schuurmans,Sergey Levine,Pieter Abbeel

Building scalable models to learn from diverse, multimodal data remains an open challenge.

For vision-language data, the dominant approaches are based on contrastive learning objectives that train a separate encoder for each modality. While effective, contrastive learning approaches introduce sampling bias depending on the data augmentations used, which can degrade performance on downstream tasks. Moreover, these methods are limited to paired image-text data, and cannot leverage widely-available unpaired data. In this paper, we investigate whether a large multimodal model trained purely via masked token prediction, without using modality-specific encoders or contrastive learning, can learn transferable representations for downstream tasks. We propose a simple and scalable network architecture, the Multimodal Masked Autoencoder (M3AE), which learns a unified encoder for both vision and language data via masked token prediction. We provide an empirical study of M3AE trained on a large-scale image-text dataset, and find that M3AE is able to learn generalizable representations that transfer well to downstream tasks. Surprisingly, we find that M3AE benefits from a higher text mask ratio (50-90%), in contrast to BERT whose standard masking ratio is 15%, due to the joint training of two data modalities. We also provide qualitative analysis showing that the learned representation incorporates meaningful information from both image and language. Lastly, we demonstrate the scalability of M3AE with larger model size and training time, and its flexibility to train on both paired image-text data as well as unpaired data.

****

Adversarial Causal Augmentation for Graph Covariate Shift

Yongduo Sui,Xiang Wang,Jiancan Wu,An Zhang,Xiangnan He,Tat-Seng Chua

Out-of-distribution (OOD) generalization on graphs is drawing widespread attention. However, existing efforts mainly focus on the OOD issue of correlation shift. While another type, covariate shift, remains largely unexplored but is the focus of this work. From a data generation view, causal features are stable substructures in data, which play key roles in OOD generalization. While their complementary parts, environments, are unstable features that often lead to various distribution shifts. Correlation shift establishes spurious statistical correlations between environments and labels. In contrast, covariate shift means that there exist unseen environmental features in test data. Existing strategies of graph invariant learning and data augmentation suffer from limited environments or unstable causal features, which greatly limits their generalization ability on covariate shift. In view of that, we propose a novel graph augmentation strategy: Adversarial Causal Augmentation (AdvCA), to alleviate the covariate shift. Specifically, it adversarially augments the data to explore diverse distributions of the environments. Meanwhile, it keeps the causal features invariant across diverse environments. It maintains the environmental diversity while ensuring the invariance of the causal features, thereby effectively alleviating the covariate shift. Extensive experimental results with in-depth analyses demonstrate that AdvCA can outperform 14 baselines on synthetic and real-world datasets with various covariate shifts.

****

Learning from conflicting data with hidden contexts

Tianren Zhang,Yizhou Jiang,Xin Su,Shangqi Guo,Chongkai Gao,Feng Chen

Classical supervised learning assumes a stable relation between inputs and outputs. However, this assumption is often invalid in real-world scenarios where the input-output relation in the data depends on some hidden contexts. We formulate a more general setting where the training data is sampled from multiple unobservable domains, while different domains may possess semantically distinct input-output maps. Training data exhibits inherent conflict in this setting, rendering vanilla empirical risk minimization problematic. We propose to tackle this problem by introducing an allocation function that learns to allocate conflicting data to different prediction models, resulting in an algorithm that we term LEAF. We draw an intriguing connection between our approach and a variant of the Expectation-Maximization algorithm. We provide theoretical justifications for LEAF on its identifiability, learnability, and generalization error. Empirical results demonstrate the efficacy and potential applications of LEAF in a range of regression and classification tasks on both synthetic data and real-world datasets.

****

Building a Subspace of Policies for Scalable Continual Learning

Jean-Baptiste Gaya,Thang Doan,Lucas Caccia,Laure Soulier,Ludovic Denoyer,Roberta Raileanu

The ability to continuously acquire new knowledge and skills is crucial for autonomous agents. Existing methods are typically based on either fixed-size models that struggle to learn a large number of diverse behaviors, or growing-size models that scale poorly with the number of tasks. In this work, we aim to strike a better balance between scalability and performance by designing a method whose size grows adaptively depending on the task sequence. We introduce Continual Subspace of Policies (CSP), a new approach that incrementally builds a subspace of policies for training a reinforcement learning agent on a sequence of tasks. The subspace's high expressivity allows CSP to perform well for many different tasks while growing more slowly than the number of tasks. Our method does not suffer from forgetting and also displays positive transfer to new tasks. CSP outperforms a number of popular baselines on a wide range of scenarios from two challenging domains, Brax (locomotion) and Continual World (robotic manipulation). Interactive visualizations of the subspace can be found at https://share.streamlit.io/continual-subspace/policies/main.

****

Complexity-Based Prompting for Multi-step Reasoning

Yao Fu,Hao Peng,Ashish Sabharwal,Peter Clark,Tushar Khot

We study the task of prompting large-scale language models to perform multi-step reasoning. Existing work shows that when prompted with a chain of thoughts (CoT), sequences of short sentences describing intermediate reasoning steps towards a final answer, large language models can generate new reasoning chains and predict answers for new inputs. A central question is which reasoning examples make the most effective prompts. In this work, we propose complexity-based prompting, a simple and effective example selection scheme for multi-step reasoning. We show that prompts with higher reasoning complexity, i.e., chains with more reasoning steps, achieve substantially better performance on math word reasoning tasks over strong baselines. We further extend our complexity-based criteria from prompting (selecting inputs) to decoding (selecting outputs), where we sample multiple reasoning chains from the model, then choose the majority of generated answers from complex reasoning chains (over simple chains). When used to prompt GPT-3, our approach substantially improves multi-step reasoning accuracy, with an 8.6% absolute improvement on GSM8K, and 6.4% on MathQA. Compared with existing example selection schemes like manual tuning or retrieval-based selection, selection based on reasoning complexity is intuitive, easy to implement, and annotation-efficient. Further results demonstrate the robustness of performance gains from complex prompts under format perturbation and distribution shift.

****

Not All Tasks Are Born Equal: Understanding Zero-Shot Generalization

Jing Zhou,Zongyu Lin,Yanan Zheng,Jian Li,Zhilin Yang

Recent work has achieved remarkable zero-shot performance with multi-task prompted pretraining, but little has been understood.
For the first time, we show that training on a small number of key tasks beats using all the training tasks, while removing these key tasks substantially hurts performance. We also find that these key tasks are mostly question answering (QA) tasks.
These novel findings combined deepen our understanding about zero-shot generalization---training on certain tasks such as QA encodes general knowledge transferable to a wide range of tasks.
In addition, to automate this procedure, we devise a method to identify and upsample key training tasks without observing the test tasks based on cross validation. Empirically, our approach achieves improved results across various model scales and tasks.

****

MA2QL: A Minimalist Approach to Fully Decentralized Multi-Agent Reinforcement Learning

Kefan Su,Siyuan Zhou,Chuang Gan,Xiangjun Wang,Zongqing Lu

Decentralized learning has shown great promise for cooperative multi-agent reinforcement learning (MARL). However, non-stationarity remains a significant challenge in fully decentralized learning. In the paper, we tackle the non-stationarity problem in the simplest and fundamental way and propose multi-agent alternate Q-learning (MA2QL), where agents take turns to update their Q-functions by Q-learning. MA2QL is a minimalist approach to fully decentralized cooperative MARL but is theoretically grounded. We prove that when each agent guarantees ε-convergence at each turn, their joint policy converges to a Nash equilibrium. In practice, MA2QL only requires minimal changes to independent Q-learning (IQL). We empirically evaluate MA2QL on a variety of cooperative multi-agent tasks. Results show MA2QL consistently outperforms IQL, which verifies the effectiveness of MA2QL, despite such minimal changes.

****

Representation Interference Suppression via Non-linear Value Factorization for Indecomposable Markov Games

Lipeng Wan,Xu He,Zeyang Liu,Kai Li,Mengchen Zhao,Dong Li,Bo An,Jianye HAO,Xuguang Lan

Value factorization is an efficient approach for centralized training with decentralized execution in cooperative multi-agent reinforcement learning tasks. As the simplest implementation of value factorization, Linear Value Factorization (LVF) attracts wide attention. In this paper, firstly, we investigate the applicable conditions of LVF, which is important but usually neglected by previous works. We prove that due to the representation limitation, LVF is only perfectly applicable to an extremely narrow class of tasks, which we define as the decomposable Markov game. Secondly, to handle the indecomposable Markov game where the LVF is inapplicable, we turn to value factorization with complete representation capability (CRC) and explore the general form of the value factorization function that satisfies both Independent Global Max (IGM) and CRC conditions. A common problem of these value factorization functions is the representation interference among true Q values with shared local Q value functions. As a result, the policy could be trapped in local optimums due to the representation interference on the optimal true Q values. Thirdly, to address the problem, we propose a novel value factorization method, namely Q Factorization with Representation Interference Suppression (QFRIS). QFRIS adaptively reduces the gradients of the local Q value functions contributed by the non-optimal true Q values. Our method is evaluated on various benchmarks. Experimental results demonstrate the good convergence of QFIRS.

****

SDAC: Efficient Safe Reinforcement Learning with Low-Biased Distributional Actor-Critic

Dohyeong Kim,Kyungjae Lee,Songhwai Oh

To apply reinforcement learning (RL) to real-world practical applications, agent

s are required to adhere to the safety guidelines of their respective domains.
Safe RL can effectively handle the guidelines by maximizing returns while maintaining safety satisfaction.
In this paper, we develop a safe distributional RL method based on the trust region method which has the capability of satisfying safety constraints consistently.
However, importance sampling required for the trust region method can hinder performance due to its significant variance, and policies may not meet the safety guidelines due to the estimation bias of distributional critics.
Hence, we enhance safety performance through the following approaches.
First, we propose novel surrogates for the trust region method expressed with Q-functions using the reparameterization trick.
Second, we utilize distributional critics trained with a target distribution where bias-variance can be traded off.
In addition, if an initial policy violates safety constraints, there can be no policy satisfying safety constraints within the trust region.
Thus, we propose a gradient integration method which is guaranteed to find a policy satisfying multiple constraints from an unsafe initial policy.
From extensive experiments, the proposed method shows minimal constraint violations while achieving high returns compared to existing safe RL methods.
Furthermore, we demonstrate the benefit of safe RL for problems in which the reward function cannot be easily specified.
****
So-TVAE: Sentiment-oriented Transformer-based Variational Autoencoder Network for Live Video Commenting
Fengyi Fu,Shancheng Fang,Weidong Chen,Yan Song,Zhendong Mao,Yongdong Zhang
Automatic live video commenting is with increasing attention due to its significance in narration generation, topic explanation, etc. However, the sentiment consideration of the generated comments is missing from the current methods. Thus, in this paper, we introduce and investigate a task, namely sentiment-guided automatic live video commenting, which aims to generate live video comments based on sentiment guidance. To address this problem, we propose a Sentiment-oriented Transformer-based Variational Autoencoder (So-TVAE) network, which consists of a sentiment-oriented diversity encoder module and a batch-attention module. Specifically, our sentiment-oriented diversity encoder elegantly combines VAE and random mask mechanism to achieve semantic diversity under sentiment guidance, which is then fused with cross-modal features to generate live video comments. Furthermore, a batch attention module is also proposed in this paper to alleviate the problem of missing sentimental samples, caused by the data imbalance, which is common in live videos as the popularity of video varies. Extensive experiments on Livebot and VideoIC datasets demonstrate that the proposed So-TVAE outperforms the state-of-the-art methods in terms of the quality and diversity of generated comments. Related codes will be released.
****
SoTeacher: Toward Student-oriented Teacher Network Training for Knowledge Distillation
Chengyu Dong,Liyuan Liu,Jingbo Shang
How to train an ideal teacher 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 practice in teacher training and the distillation objective. To fill this gap, we explore the feasibility of training a teacher that is oriented toward student performance with empirical risk minimization. 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 established that (1) the empirical risk minimizer can provably approximate the true label distribution of training data if the loss function is a proper scoring rule and the hypothesis function is locally-Lipschitz continuous around training inputs; and (2) when data augmentation is employed for training, an additional constraint is required that the minimizer has

to produce consistent predictions across augmented views of the same training input. In light of our theory, we propose a teacher training method SoTeacher which renovates the empirical risk minimization by incorporating Lipschitz regularization and consistency regularization. Experiments on two benchmark datasets confirm that SoTeacher can improve student performance significantly and consistently across various knowledge distillation algorithms and teacher-student pairs.

****

GuardHFL: Privacy Guardian for Heterogeneous Federated Learning

Hanxiao Chen,Meng Hao,Hongwei Li,Guangxiao Niu,Guowen Xu,Tianwei Zhang,Xilin Zhang

Heterogeneous federated learning (HFL) enables clients with different computation and communication capabilities to collaboratively train their own customized models via a query-response paradigm on auxiliary datasets. However, such paradigm raises serious privacy issues due to the leakage of highly sensitive query samples and response predictions. Although existing secure querying solutions may be extended to enhance the privacy of HFL with non-trivial adaptation, they suffer from two key limitations: (1) lacking customized protocol designs and (2) relying on heavy cryptographic primitives, which could lead to poor performance. In this work, we put forth GuardHFL, the first-of-its-kind efficient and privacy-preserving HFL framework. GuardHFL is equipped with a novel HFL-friendly secure querying scheme that is built on lightweight secret sharing and symmetric-key techniques. Its core is a set of customized multiplication and comparison protocols, which substantially boost the execution efficiency. Extensive evaluations demonstrate that GuardHFL outperforms the state-of-the-art works by up to two orders of magnitude in efficiency.

****

Decentralized Policy Optimization

Kefan Su,Zongqing Lu

The study of decentralized learning or independent learning in cooperative multi-agent reinforcement learning has a history of decades. Recently empirical studies show that independent PPO (IPPO) can obtain good performance, close to or even better than the methods of centralized training with decentralized execution, in several benchmarks. However, decentralized actor-critic with convergence guarantee is still open. In this paper, we propose decentralized policy optimization (DPO), a decentralized actor-critic algorithm with monotonic improvement and convergence guarantee. We derive a novel decentralized surrogate for policy optimization such that the monotonic improvement of joint policy can be guaranteed by each agent independently optimizing the surrogate. In practice, this decentralized surrogate can be realized by two adaptive coefficients for policy optimization at each agent. Empirically, we compare DPO with IPPO in a variety of cooperative multi-agent tasks, covering discrete and continuous action spaces, and fully and partially observable environments. The results show DPO outperforms IPPO in most tasks, which can be the evidence for our theoretical results.

****

Identification of the Adversary from a Single Adversarial Example

Minhao Cheng,Rui Min,Haochen Sun

Deep neural networks have been shown vulnerable to adversarial examples. Even though many defence methods have been proposed to enhance the robustness, it is still a long way toward providing an attack-free method to build a trustworthy machine learning system. In this paper, instead of enhancing the robustness, we take the investigator's perspective and propose a new framework to trace the first compromised model in a forensic investigation manner. Specifically, we focus on the following setting: the machine learning service provider provides models for a set of customers. However, one of the customers conducted adversarial attacks to fool the system. Therefore, the investigator's objective is to identify the first compromised model by collecting and analyzing evidence from only available adversarial examples. To make the tracing viable, we design a random mask watermarking mechanism to differentiate adversarial examples from different models. First, we propose a tracing approach in the data-limited case where the original example is also available. Then, we design a data-free approach to identify the adversary without accessing the original example. Finally, the effectiveness of our proposed framework is evaluated by extensive experiments with different model architectures, adversarial attacks, and datasets.

****

Similarity of Neural Architectures Based on Input Gradient Transferability
Jaehui Hwang,Dongyoon Han,Byeongho Heo,Song Park,Sanghyuk Chun,Jong-Seok Lee

In this paper, we aim to design a quantitative similarity function between two neural architectures. Specifically, we define a model similarity using input gradient transferability. We generate adversarial samples of two networks and measure the average accuracy of the networks on adversarial samples of each other. If two networks are highly correlated, then the attack transferability will be high, resulting in high similarity. Using the similarity score, we investigate two topics: (1) Which network component contributes to the model diversity? (2) How does model diversity affect practical scenarios? We answer the first question by providing feature importance analysis and clustering analysis. The second question is validated by two different scenarios: model ensemble and knowledge distillation. Our findings show that model diversity takes a key role when interacting with different neural architectures. For example, we found that more diversity leads to better ensemble performance. We also observe that the relationship between teacher and student networks and distillation performance depends on the choice of the base architecture of the teacher and student networks. We expect our analysis tool helps a high-level understanding of differences between various neural architectures as well as practical guidance when using multiple architectures.

****

Image Segmentation using Transfer Learning with DeepLabv3 to Facilitate Photogrammetric Limb Scanning
Isaac A Cabrera,Yixuan Zhou,Eric K Ngo,Ramesh Rao,Albert Lin

In this paper, we explore the use of deep learning (DL) in conjunction with photogrammetry for scanning amputated limbs. Combining these two technologies can expand the scope of prosthetic telemedicine by facilitating low-cost limb scanning using cell phones. Previous research identified image segmentation as one of the main limitations of using photogrammetry for limb scanning. Based on those limitations, this work sought to answer two main research questions: (1) Can a neural network be trained to identify and segment an amputated limb automatically? (2) Will segmenting 2D limb images using neural networks impact the accuracy of 3D models generated via photogrammetry? To answer the first question, transfer learning was applied to a neural network with the DeepLabv3 architecture. After training, the model was able to successfully identify and segment limb images with an IoU of 79.9%. To answer the second question, the fine-tuned DL model was applied to a dataset of 22 scans comprising 6312 limb images, then 3D models were rendered utilizing Agisoft Metashape. The Mean Absolute Error (MAE) of models rendered from images segmented with DL was 0.57 mm ± 0.63 mm when compared to models rendered from ground truth images. These results are important because segmentation with DL makes photogrammetry for limb scanning feasible on a large clinical scale. Future work should focus on generalizing the segmentation model for different types of amputations and imaging conditions.

****

G-Censor: Graph Contrastive Learning with Task-Oriented Counterfactual Views
tianqianjin lin,Yangyang Kang,Zhuoren Jiang,Xurui Li,Changlong Sun,cui huang,Xiaozhong Liu

Graph Contrastive learning (GCL) has achieved great success in learning representations from unlabeled graph-structure data. However, how to automatically obtain the optimal contrastive views w.r.t specific downstream tasks is little studied. Theoretically, a downstream task can be causally correlated to particular sub-structures in graphs. The existing GCL methods may fail to enhance model performance on a given task when the task-related semantics are incomplete/preserved in the positive/negative views. To address this problem, we propose G-CENSOR, i.e., Graph Contrastive lEarniNg with taSk-oriented cOunteRfactual views, a model-agnostic framework designed for node property prediction tasks. G-CENSOR can simu

ltaneously generate the optimal task-oriented counterfactual positive/negative views for raw ego-graphs and train graph neural networks (GNNs) with a contrastive objective between the raw ego-graphs and their corresponding counterfac-tual views. Extensive experiments on eight real-world datasets demonstrate that G-CENSOR can consistently outperform existing state-of-the-art GCL methods to improve the task performance and generalizability of a series of typical GNNs. To the best of our knowledge, this is a pioneer investigation to explore task-oriented graph contrastive learning from a counterfactual perspective in node property prediction tasks. We will release the source code after the review process.
****

Unsupervised 3D Object Learning through Neuron Activity aware Plasticity
Beomseok Kang,Biswadeep Chakraborty,Saibal Mukhopadhyay

We present an unsupervised deep learning model for 3D object classification. Conventional Hebbian learning, a well-known unsupervised model, suffers from loss of local features leading to reduced performance for tasks with complex geometric objects. We present a deep network with a novel Neuron Activity Aware (NeAW) Hebbian learning rule that dynamically switches the neurons to be governed by Hebbian learning or anti-Hebbian learning, depending on its activity. We analytically show that NeAW Hebbian learning relieves the bias in neuron activity, allowing more neurons to attend to the representation of the 3D objects. Empirical results show that the NeAW Hebbian learning outperforms other variants of Hebbian learning and shows higher accuracy over fully supervised models when training data is limited.
****

Visually-Augmented Language Modeling
Weizhi Wang,Li Dong,Hao Cheng,Haoyu Song,Xiaodong Liu,Xifeng Yan,Jianfeng Gao,Furu Wei

Human language is grounded on multimodal knowledge including visual knowledge like colors, sizes, and shapes. However, current large-scale pre-trained language models rely on the text-only self-supervised training with massive text data, which precludes them from utilizing relevant visual information when necessary. To address this, we propose a novel pre-training framework, named VaLM, to Visually-augment text tokens with retrieved relevant images for Language Modeling. Specifically, VaLM builds on a novel latent text-image alignment method via an image retrieval module to fetch corresponding images given a textual context. With the visually-augmented context, VaLM uses a visual knowledge fusion layer to enable multimodal grounded language modeling by attending on both text context and visual knowledge in images. We evaluate VaLM on various visual knowledge intensive commonsense reasoning tasks, which require visual information to excel. The experimental results illustrate that VaLM outperforms all strong language-only and vision-language baselines with substantial gains on reasoning object commonsense including color, size, and shape.
****

Multi-Layered 3D Garments Animation
Yidi Shao,Chen Change Loy,Bo Dai

Most existing 3D garment animation datasets are restricted to human bodies with single-layered garments. Even though cases with upper shirts and lower pants are included, only a few overlap areas among such garment combinations exist. Moreover, they often regard human body movement as the only driving factor that causes garment animation. Approaches developed on top of these datasets thus tend to model garments as functions of human body parameters such as body shape and pose. While such treatment leads to promising performance on existing datasets, it leaves a gap between experimental environments and real scenarios, where a body can wear multiple layered garments and the corresponding garment dynamics can be affected by environmental factors and garment attributes. Consequently, existing approaches often struggle to generalize to multi-layered garments and realistic scenarios. To facilitate the advance of 3D garment animation toward handling more challenging cases, this paper presents a new large-scale synthetic dataset called LAYERS, covering 4,900 different combinations of multi-layered garments with 700k frames in total. The animation of these multi-layered garments follows th e laws of physics and is affected by not only human body movements but also random environmental wind and garment attributes. To demonstrate the quality of LAYERS, we further propose a novel method, LayersNet, for 3D garment animation, which represents garments as unions of particles and subsequently adopts a neural network to animate garments via particle-based simulation. In this way, the interactions between different parts of one garment, different garments on the same body, and garments against various driving factors, can be naturally and uniformly handled via the interactions of particles. Through comprehensive experiments, LayersNet demonstrates superior performance in terms of animation accuracy and generality over baselines. The proposed dataset, LAYERS, as well as the proposed method, LayersNet, will be publicly available.

****

Unsupervised Learning of Structured Representations via Closed-Loop Transcription

Shengbang Tong,Xili Dai,Yubei Chen,Mingyang Li,ZENGYI LI,Brent Yi,Yann LeCun,Yi Ma

This paper proposes an unsupervised method for learning a unified representation that serves both discriminative and generative purposes. While most existing unsupervised learning approaches focus on a representation for only one of these two goals, we show that a unified representation can enjoy the mutual benefits of having both. Such a representation is attainable by generalizing the recently proposed closed-loop transcription framework, known as CTRL, to the unsupervised setting. This entails solving a constrained maximin game over a rate reduction objective that expands features of all samples while compressing features of augmentations of each sample. Through this process, we see discriminative low-dimensional structures emerge in the resulting representations. Under comparable experimental conditions and network complexities, we demonstrate that these structured representations enable classification performance close to state-of-the-art unsupervised discriminative representations, and conditionally generated image quality significantly higher than that of state-of-the-art unsupervised generative models.

****

Closed Boundary Learning for NLP Classification Tasks with the Universum Class

Hanzhang Zhou,Zijian Feng,Kezhi Mao,Edmond Y.M. Lo,Lihui Chen

The Universum class, often known as the other class or the miscellaneous class, is defined as a collection of samples that do not belong to any class of interest. It is a typical class that exists in many classification-based tasks in natural language processing (NLP), such as relation extraction, named entity recognition, sentiment analysis, etc. During data labeling, a significant number of samples are annotated as Universum because there are always some samples that exist in the dataset but do not belong to preset target classes and are not of interest in the task. The Universum class exhibits very different properties, namely heterogeneity and lack of representativeness in training data; however, existing methods often treat the Universum class equally with the classes of interest. Although the Universum class only contains uninterested samples, improper treatment will result in the misclassification of samples of interest. In this work, we propose a closed boundary learning method that treats the Universum class and classes of interest differently. We apply closed decision boundaries to classes of interest and designate the area outside all closed boundaries in the feature space as the space of the Universum class. Specifically, we formulate the closed boundaries as arbitrary shapes, propose a strategy to estimate the probability of the Universum class according to its unique property rather than the within-class sample distribution, and propose a boundary learning loss to learn decision boundaries based on the balance of misclassified samples inside and outside the boundary. We evaluate our method on 6 state-of-the-art works in 3 different tasks, and the performance of all 6 works is improved. Our code will be released on GitHub.

****

Solving Constrained Variational Inequalities via a First-order Interior Point-based Method

Tong Yang,Michael Jordan,Tatjana Chavdarova

We develop an interior-point approach to solve constrained variational inequality (cVI) problems. Inspired by the efficacy of the alternating direction method of multipliers (ADMM) method in the single-objective context, we generalize ADMM to derive a first-order method for cVIs, that we refer to as ADMM-based interior-point method for constrained VIs (ACVI). We provide convergence guarantees for ACVI in two general classes of problems: (i) when the operator is ξ-monotone, and (ii) when it is monotone, some constraints are active and the game is not purely rotational. When the operator is in addition L-Lipschitz for the latter case, we match known lower bounds on rates for the gap function of $\mathcal{O}(1/\sqrt{K})$ and $\mathcal{O}(1/K)$ for the last and average iterate, respectively. To the best of our knowledge, this is the first presentation of a first-order interior-point method for the general cVI problem that has a global convergence guarantee. Moreover, unlike previous work in this setting, ACVI provides a means to solve cVIs when the constraints are nontrivial. Empirical analyses demonstrate clear advantages of ACVI over common first-order methods. In particular, (i) cyclical behavior is notably reduced as our methods approach the solution from the analytic center, and (ii) unlike projection-based methods that zigzag when near a constraint, ACVI efficiently handles the constraints.

****

MeGraph: Graph Representation Learning on Connected Multi-scale Graphs

Honghua Dong,Jiawei Xu,Yu Yang,Rui Zhao,Chun Yuan,Xiu Li,Chris J. Maddison,Lei Han

We present MeGraph, a novel network architecture for graph-structured data. Given any input graph, we create multi-scale graphs using graph pooling. Then, we connect them into a mega graph by bridging inter-graph edges according to the graph pooling results. Instead of universally stacking graph convolutions over the mega graph, we apply general graph convolutions over intra-graph edges, while the convolutions over inter-graph edges follow a bidirectional pathway to deliver the information along the hierarchy for one turn. Graph convolution and graph pooling are two core elementary operations of MeGraph. In our implementation, we adopt the graph full network (GFuN) and propose the stridden edge contraction pooling (S-EdgePool) with adjustable pooling ratio, which are extended from conventional graph convolution and edge contraction pooling. The MeGraph model enables information exchange across multi-scale graphs, repeatedly, for deeper understanding of wide-range correlations in graphs. This distinguishes MeGraph from many recent hierarchical graph neural networks like Graph U-Nets. We conduct comprehensive empirical studies on tens of public datasets, in which we observe consistent performance gains comparing to baselines. Specifically, we establish 5 new graph theory benchmark tasks that require long-term inference and deduction to solve, where MeGraph demonstrates dominated performance compared with popular graph neural networks.

****

Learning Reduced Fluid Dynamics

zherong pan,Xifeng Gao,Kui Wu

Predicting the state evolution of ultra high-dimensional, time-reversible fluid dynamic system is a crucial but computationally expensive task. Model-reduction has been proven to be an effective method to reduce the computational cost by learning a low-dimensional state embedding. However, existing reduced models are irrespective of either the time reversible property or the nonlinear dynamics, leading to sub-optimal performance. We propose a model-based approach to identify locally optimal, model-reduced, time reversible, nonlinear fluid dynamic systems. Our main idea is to use stochastic Riemann optimization to obtain a high-quality a reduced fluid model by minimizing the expected trajectory-wise model-reduction error over a given distribution of initial conditions. To this end, our method formulates the reduced fluid dynamics as an invertible state transfer function parameterized by the reduced subspace. We further show that the reduced trajectories are differentiable with respect to the subspace bases over the entire Grassmannian manifold, under proper choices of timestep sizes and numerical integrators. Finally, we propose a loss function measuring the trajectory-wise discrepa ncy between the original and reduced models. By tensor precomputation, we show that gradient information of such loss functions can be evaluated efficiently over a long trajectory without time-integrating the high-dimensional dynamic system. Through evaluations on a row of simulation benchmarks, we show that our method lower the discrepancy by 45%-97% over conventional reduced models.

****

Symmetric Pruning in Quantum Neural Networks

Xinbiao Wang,Junyu Liu,Tongliang Liu,Yong Luo,Yuxuan Du,Dacheng Tao

Many fundamental properties of a quantum system are captured by its Hamiltonian and ground state. Despite the significance, ground states preparation (GSP) is classically intractable for large-scale Hamiltonians. Quantum neural networks (QNNs), which exert the power of modern quantum machines, have emerged as a leading protocol to conquer this issue. As such, the performance enhancement of QNNs becomes the core in GSP. Empirical evidence showed that QNNs with handcraft symmetric ans\"atze generally experience better trainability than those with asymmetric ans\"atze, while theoretical explanations remain vague. To fill this knowledge gap, here we propose the effective quantum neural tangent kernel (EQNTK) and connect this concept with over-parameterization theory to quantify the convergence of QNNs towards the global optima. We uncover that the advance of symmetric ans\"atze attributes to their large EQNTK value with low effective dimension, which requests few parameters and quantum circuit depth to reach the over-parameterization regime permitting a benign loss landscape and fast convergence. Guided by EQNTK, we further devise a symmetric pruning (SP) scheme to automatically tailor a symmetric ansatz from an over-parameterized and asymmetric one to greatly improve the performance of QNNs when the explicit symmetry information of Hamiltonian is unavailable. Extensive numerical simulations are conducted to validate the analytical results of EQNTK and the effectiveness of SP.

****

Managing Temporal Resolution in Continuous Value Estimation: A Fundamental Trade-off

Zichen Zhang,Johannes Kirschner,Junxi Zhang,Francesco Zanini,Alex Ayoub,Masood Dehghan,Dale Schuurmans

A default assumption in reinforcement learning and optimal control is that experience arrives at discrete time points on a fixed clock cycle. Many applications, however, involve continuous systems where the time discretization is not fixed but instead can be managed by a learning algorithm. By analyzing Monte-Carlo value estimation for LQR systems in both finite-horizon and infinite-horizon settings, we uncover a fundamental trade-off between approximation and statistical error in value estimation. Importantly, these two errors behave differently with respect to time discretization, which implies that there is an optimal choice for the temporal resolution that depends on the data budget. These findings show how adapting the temporal resolution can provably improve value estimation quality in LQR systems from finite data. Empirically, we demonstrate the trade-off in numerical simulations of LQR instances and several non-linear environments.

****

On the Robustness of Randomized Ensembles to Adversarial Perturbations

Hassan Dbouk,Naresh Shanbhag

Randomized ensemble classifiers (RECs), where one classifier is randomly selected during inference, have emerged as an attractive alternative to traditional ensembling methods for realizing adversarially robust classifiers with limited compute requirements. However, recent works have shown that existing methods for constructing RECs are more vulnerable than initially claimed, casting major doubts on their efficacy and prompting fundamental questions such as: "When are RECs useful?", "What are their limits?", and "How do we train them?". In this work, we first demystify RECs as we derive fundamental results regarding their theoretical limits, necessary and sufficient conditions for them to be useful, and more. Leveraging this new understanding, we propose a new boosting
algorithm (BARRE) for training robust RECs, and empirically demonstrate its effectiveness at defending against strong ℓ_∞ norm-bounded adversaries across various network architectures and datasets. Our code is submitted as part of

the supplementary material, and will be publicly released on GitHub

****

Minimum Variance Unbiased N:M Sparsity for the Neural Gradients

Brian Chmiel,Itay Hubara,Ron Banner,Daniel Soudry

In deep learning, fine-grained N:M sparsity reduces the data footprint and bandwidth of a General Matrix multiply (GEMM) up to x2, and doubles throughput by skipping computation of zero values. So far, it was mainly only used to prune weights to accelerate the forward and backward phases. We examine how this method can be used also for the neural gradients (i.e. loss gradients with respect to the intermediate neural layer outputs). To this end, we first establish a tensor-level optimality criteria. Previous works aimed to minimize the mean-square-error (MSE) of each pruned block. We show that while minimization of the MSE works fine for pruning the weights and activations, it catastrophically fails for the neural gradients. Instead, we show that accurate pruning of the neural gradients requires an unbiased minimum-variance pruning mask. We design such specialized masks, and find that in most cases, 1:2 sparsity is sufficient for training, and 2:4 sparsity is usually enough when this is not the case. Further, we suggest combining several such methods together in order to potentially speed up training even more. A reference implementation is supplied in the supplementary material.

****

Incremental Learning of Structured Memory via Closed-Loop Transcription

Shengbang Tong,Xili Dai,Ziyang Wu,Mingyang Li,Brent Yi,Yi Ma

This work proposes a minimal computational model for learning structured memories of multiple object classes in an incremental setting. Our approach is based on establishing a {\em closed-loop transcription} between the classes and a corresponding set of subspaces, known as a linear discriminative representation, in a low-dimensional feature space. Our method is simpler than existing approaches for incremental learning, and more efficient in terms of model size, storage, and computation: it requires only a single, fixed-capacity autoencoding network with a feature space that is used for both discriminative and generative purposes. Network parameters are optimized simultaneously without architectural manipulations, by solving a constrained minimax game between the encoding and decoding maps over a single rate reduction-based objective. Experimental results show that our method can effectively alleviate catastrophic forgetting, achieving significantly better performance than prior work of generative replay on MNIST, CIFAR-10, and ImageNet-50, despite requiring fewer resources.

****

Curved Data Representations in Deep Learning

Ilya Kaufman,Omri Azencot

The phenomenal success of deep neural networks inspire many to understand the inner mechanisms of these models. To this end, several works have been studying geometric properties such as the intrinsic dimension of latent data representations produced by the layers of the network. In this paper, we investigate the curvature of data manifolds, i.e., the deviation of the manifold from being flat in its principal directions. We find that state-of-the-art trained convolutional neural networks have a characteristic curvature profile along layers: an initial increase, followed by a long phase of a plateau, and tailed by another increase. In contrast, untrained networks exhibit qualitatively and quantitatively different curvature profiles. We also show that the curvature gap between the last two layers is strongly correlated with the performance of the network. Further, we find that the intrinsic dimension of latent data along the network layers is not necessarily indicative of curvature. Finally, we evaluate the effect of common regularizers such as weight decay and mixup on curvature, and we find that mixup-based methods flatten intermediate layers, whereas the final layers still feature high curvatures. Our results indicate that relatively flat manifolds which transform to highly-curved manifolds toward the last layers generalize well to unseen data.

****

When Data Geometry Meets Deep Function: Generalizing Offline Reinforcement Learning

Jianxiong Li,Xianyuan Zhan,Haoran Xu,Xiangyu Zhu,Jingjing Liu,Ya-Qin Zhang
In offline reinforcement learning (RL), one detrimental issue to policy learning is the error accumulation of deep \textit{Q} function in out-of-distribution (OOD) areas. Unfortunately, existing offline RL methods are often over-conservative, inevitably hurting generalization performance outside data distribution. In our study, one interesting observation is that deep \textit{Q} functions approximate well inside the convex hull of training data. Inspired by this, we propose a new method, \textit{DOGE (Distance-sensitive Offline RL with better GEneralization)}. DOGE marries dataset geometry with deep function approximators in offline RL, and enables exploitation in generalizable OOD areas rather than strictly constraining policy within data distribution. Specifically, DOGE trains a state-conditioned distance function that can be readily plugged into standard actor-critic methods as a policy constraint. Simple yet elegant, our algorithm enjoys better generalization compared to state-of-the-art methods on D4RL benchmarks. Theoretical analysis demonstrates the superiority of our approach to existing methods that are solely based on data distribution or support constraints.
****

Self-supervised debiasing using low rank regularization
Geon Yeong Park,Chanyong Jung,Jong Chul Ye,Sang Wan Lee
Spurious correlations can cause strong biases in deep neural networks, impairing generalization ability. While most of existing debiasing methods require full supervisions on either spurious attributes or target labels, training a debiased model from a limited amount of both annotations is still an open issue. To overcome such limitations, we first examined an interesting phenomenon by the spectral analysis of latent representations: spuriously correlated, easy-to-learn attributes make neural networks inductively biased towards encoding lower effective rank representations. We also show that a rank regularization can amplify this bias in a way that encourages highly correlated features. Motivated by these observations, we propose a self-supervised debiasing framework that is potentially compatible with unlabeled samples. We first pretrain a biased encoder in a self-supervised manner with the rank regularization, serving as a semantic bottleneck to enforce the encoder to learn the spuriously correlated attributes. This biased encoder is then used to discover and upweight bias-conflicting samples in a downstream task, serving as a boosting to effectively debias the main model. Remarkably, the proposed debiasing framework significantly improves the generalization performance of self-supervised learning baselines and, in some cases, even outperforms state-of-the-art supervised debiasing approaches.
****

Wasserstein Gradient Flows for Optimizing GMM-based Policies
Hanna Ziesche,Leonel Rozo
Robots often rely on a repertoire of previously-learned motion policies for performing tasks of diverse complexities.
When facing unseen task conditions or when new task requirements arise, robots must adapt their motion policies accordingly.
In this context, policy optimization is the de facto paradigm to adapt robot policies as a function of task-specific objectives.
Most commonly-used motion policies carry particular structures that are often overlooked in policy optimization algorithms.
We instead propose to leverage the structure of probabilistic policies by casting the policy optimization as an optimal transport problem.
Specifically, we focus on robot motion policies that build on Gaussian mixture models (GMMs) and formulate the policy optimization as a Wassertein gradient flow over the GMMs space.
This naturally allows us to constrain the policy updates via the L^2-Wasserstein distance between GMMs to enhance the stability of the policy optimization process.
Furthermore, we leverage the geometry of the Bures-Wasserstein manifold to optimize the Gaussian distributions of the GMM policy via Riemannian optimization.
We evaluate our approach over a set of common robotic settings: Reaching motions, collision-avoidance behaviors and multi-goal tasks.

Our results show that our method outperforms common policy optimization baselines in terms of task success rate and low-variance solutions.
****

Neural Unbalanced Optimal Transport via Cycle-Consistent Semi-Couplings
Frederike Lübeck,Charlotte Bunne,Gabriele Gut,Jacobo Sarabia del Castillo,Lucas Pelkmans,David Alvarez-Melis
Comparing unpaired samples of a distribution or population taken at different points in time is a fundamental task in many application domains where measuring populations is destructive and cannot be done repeatedly on the same sample, such as in single-cell biology. Optimal transport (OT) can solve this challenge by learning an optimal coupling of samples across distributions from unpaired data. However, the usual formulation of OT assumes conservation of mass, which is violated in unbalanced scenarios in which the population size changes (e.g., cell proliferation or death) between measurements. In this work, we introduce NubOT, a neural unbalanced OT formulation that relies on the formalism of semi-couplings to account for creation and destruction of mass. To estimate such semi-couplings and generalize out-of-sample, we derive an efficient parameterization based on neural optimal transport maps and propose a novel algorithmic scheme through a cycle-consistent training procedure. We apply our method to the challenging task of forecasting heterogeneous responses of multiple cancer cell lines to various drugs, where we observe that by accurately modeling cell proliferation and death, our method yields notable improvements over previous neural optimal transport methods.
****

Budgeted Training for Vision Transformer
zhuofan xia,Xuran Pan,Xuan Jin,Yuan He,Hui Xue',Shiji Song,Gao Huang
The superior performances of Vision Transformers often come with higher training costs. Compared to their CNN counterpart, Transformer models are hungry for large-scale data and their training schedules are usually prolonged. This sets great restrictions on training Transformers with limited resources, where a proper trade-off between training cost and model performance is longed. In this paper, we address the problem by proposing a framework that enables the training process under \textit{any training budget} from the perspective of model structure, while achieving competitive model performances. Specifically, based on the observation that Transformer exhibits different levels of model redundancies at different training stages, we propose to dynamically control the activation rate of the model structure along the training process and meet the demand on the training budget by adjusting the duration on each level of model complexity. Extensive experiments demonstrate that our framework is applicable to various Vision Transformers, and achieves competitive performances on a wide range of training budgets.

****

Knowledge-in-Context: Towards Knowledgeable Semi-Parametric Language Models
Xiaoman Pan,Wenlin Yao,Hongming Zhang,Dian Yu,Dong Yu,Jianshu Chen
Fully-parametric language models generally require a huge number of model parameters to store the necessary knowledge for solving multiple natural language tasks in zero/few-shot settings. In addition, it is hard to adapt to the evolving world knowledge without the costly model re-training. In this paper, we develop a novel semi-parametric language model architecture, Knowledge-in-Context (KiC), which empowers a parametric text-to-text language model with a knowledge-rich external memory. Specifically, the external memory contains six different types of knowledge: entity, dictionary, commonsense, event, script, and causality knowledge. For each input instance, the KiC model adaptively selects a knowledge type and retrieves the most helpful pieces of knowledge. The input instance along with its knowledge augmentation is fed into a text-to-text model (e.g., T5) to generate the output answer, where both the input and the output are in natural language forms after prompting. Interestingly, we find that KiC can be identified as a special mixture-of-experts (MoE) model, where the knowledge selector plays the role of a router that is used to determine the sequence-to-expert assignment in MoE. This key observation inspires us to develop a novel algorithm for trainin

g KiC with an instance-adaptive knowledge selector. As a knowledge-rich semi-parametric language model, KiC only needs a much smaller parametric part to achieve superior zero-shot performance on unseen tasks. By evaluating on 40+ different tasks, we show that KiC-Large with 770M parameters easily outperforms large language models that are 4-39x larger. In addition, KiC also exhibits emergent abilities at a much smaller model scale compared to the fully-parametric models.

****

Understanding and Mitigating Robust Overfitting through the Lens of Feature Dynamics

Yifei Wang,Liangchen Li,Yisen Wang,Jiansheng Yang,Zhouchen Lin

Adversarial Training (AT) has become arguably the state-of-the-art algorithm for extracting robust features. However, researchers recently notice that AT suffers from severe robust overfitting problems, particularly after the learning rate (LR) decay, while the existing static view of feature robustness fails to explain this phenomenon. In this paper, we propose a new dynamic feature robustness framework which takes the dynamic interplay between the model trainer and the attacker into consideration. By tracing temporal and dataset-specific feature robustness, we develop a new understanding of robust overfitting from the dynamics of non-robust features, and empirically verify it on real-world datasets. Built upon this understanding, we explore three techniques to restore the balance between the model trainer and the attacker, and show that they could effectively alleviate robust overfitting and attain state-of-the-art robustness on benchmark datasets. Notably, different from previous studies, our interpretation highlights the necessity of considering the min-max nature of AT for robust overfitting.

****

Augmentative Topology Agents For Open-Ended Learning

Muhammad Umair Nasir,Michael Beukman,Steven James,Christopher Wesley Cleghorn

In this work, we tackle the problem of Open-Ended Learning by a method that simultaneously evolves agents and increasingly challenging environments. Unlike previous open-ended approaches that optimize agents using a fixed neural network topology, we hypothesize that generalization can be improved by allowing agents' controllers to become more complex as they encounter more difficult environments. Our method, Augmentative Topology EPOET (ATEP), extends the Enhanced Paired Open-Ended Trailblazer (EPOET) algorithm by allowing agents to evolve their own neural network structures over time, adding complexity and capacity as necessary. Empirical results demonstrate that ATEP results in general agents capable of solving more environments than a fixed-topology baseline. We also investigate mechanisms for transferring agents between environments and find that a species-based approach further improves the performance and generalization of agents.

****

Learning to Boost Resilience of Complex Networks via Neural Edge Rewiring

Shanchao Yang,MA KAILI,Tianshu Yu,Baoxiang Wang,Hongyuan Zha

The resilience of complex networks, a critical structural characteristic in network science, measures the network's ability to withstand noise corruption and structural changes. Improving resilience typically resorts to minimal modifications of the network structure via degree-preserving edge rewiring-based methods. Despite their effectiveness, existing methods are learning-free, sharing the limitation of transduction: a learned edge rewiring strategy from one graph cannot be generalized to another. Such a limitation cannot be trivially addressed by existing graph neural networks (GNNs)-based approaches since there is no rich initial node features for GNNs to learn meaningful representations. However, neural edge rewiring relies on GNNs for obtaining meaningful representations from pure graph topologies to select edges. We found existing GNNs degenerate remarkably with only pure topologies on the resilience task, leading to the undesired infinite action backtracking. In this work, inspired by persistent homology, we specifically design a variant of GNN called FireGNN for learning inductive edge rewiring strategies. Based on meaningful representations from FireGNN, we develop the first end-to-end inductive method, ResiNet, to discover $\textbf{resi}$lient $\textbf{net}$work topologies while balancing network utility. ResiNet reformulates n etwork resilience optimization as a Markov decision process equipped with edge rewiring action space and learns to select correct edges successively. Extensive experiments demonstrate that ResiNet achieves a near-optimal resilience gain on various graphs while balancing the utility and outperforms existing approaches by a large margin.

****

Deep Transformer Q-Networks for Partially Observable Reinforcement Learning
Kevin Esslinger,Robert Platt,Christopher Amato

Real-world reinforcement learning tasks often involve some form of partial observability where the observations only give a partial or noisy view of the true state of the world. Such tasks typically require some form of memory, where the agent has access to multiple past observations, in order to perform well. One popular way to incorporate memory is by using a recurrent neural network to access the agent's history. However, recurrent neural networks in reinforcement learning are often fragile and difficult to train, susceptible to catastrophic forgetting and sometimes fail completely as a result. In this work, we propose Deep Transformer Q-Networks (DTQN), a novel architecture utilizing transformers and self-attention to encode an agent's history. DTQN is designed modularly, and we compare results against several modifications to our base model. Our experiments demonstrate the transformer can solve partially observable tasks faster and more stably than previous recurrent approaches.

****

Mind's Eye: Grounded Language Model Reasoning through Simulation
Ruibo Liu,Jason Wei,Shixiang Shane Gu,Te-Yen Wu,Soroush Vosoughi,Claire Cui,Denny Zhou,Andrew M. Dai

Successful and effective communication between humans and AI relies on a shared experience of the world. By training solely on written text, current language models (LMs) miss the grounded experience of humans in the real-world---their failure to relate language to the physical world causes knowledge to be misrepresented and obvious mistakes in their reasoning. We present Mind's Eye, a paradigm to ground language model reasoning in the physical world. Given a physical reasoning question, we use a computational physics engine (DeepMind's MuJoCo) to simulate the possible outcomes, and then use the simulation results as part of the input, which enables language models to perform reasoning. Experiments on 39 tasks in a physics alignment benchmark demonstrate that Mind's Eye can improve reasoning ability by a large margin (27.9% zero-shot, and 46.0% few-shot absolute accuracy improvement on average). Smaller language models armed with Mind's Eye can obtain similar performance to models that are 100x larger. Finally, we confirm the robustness of Mind's Eye through ablation studies.

****

Visual Expertise and the Log-Polar Transform Explain Image Inversion Effects
Martha Gahl,Shubham Kulkarni,Nikhil Pathak,Alex Russell,Garrison W. Cottrell

Visual expertise can be defined as the ability to discriminate among subordinate-level objects in homogeneous classes, such as identities of faces within the class "face". Despite being able to discriminate many faces, subjects perform poorly at recognizing even familiar faces once inverted. This face-inversion effect is in contrast to subjects’ performance identifying inverted objects for which their experience is at a basic level, which results in less impairment. Experimental results have suggested that when identifying mono-oriented objects, such as cars, car novices' performance is between that of faces and other objects. We build an anatomically-inspired neurocomputational model to explore this effect. Our model includes a foveated retina and the log-polar mapping from the visual field to V1. This transformation causes changes in scale to appear as horizontal translations, leading to scale equivariance. Rotation is similarly equivariant, leading to vertical translations. When fed into a standard convolutional network, this provides rotation and scale invariance. It may be surprising that a rotation-invariant network shows any inversion effect at all. This is because there is a crucial topological difference between scale and rotation: Rotational invariance is discontinuous, with V1 ranging from 90 degrees (vertically up) to 270 degrees (vertically down). Hence when a face is inverted, the configural informati

on in the face is disrupted while feature information is relatively unaffected. We show that the inversion effect arises as a result of visual expertise, where configural information becomes relevant as more identities are learned at the subordinate level. Our model matches the classic result: faces suffer more from inversion than mono-oriented objects, which are more disrupted than non-mono-oriented objects when objects are only familiar at a basic level.
****

Cross-Protein Wasserstein Transformer for Protein-Protein Interactions
Gongping Xu,Tong Zhang,Wenting Zhao,Zhen Cui,Jian Yang
Previous studies reveal intimate relationships between the structure and function of proteins. Motivated by this, for protein-protein interactions (PPIs), we hypothesize that cross-protein structural correspondence, including both global correlation and local co-occurrence, poses a great influence. Accordingly, a novel deep learning framework named Cross-Protein Wasserstein Transformer (CPWT) is proposed to predict PPI sites through fine-grained cross-graph structural modeling. Considering the irregular architecture of acid sequences, for a pair of proteins, graphs are constructed to describe them. Then, a core Cross-Graph Transformer (CGT) module of two branches (e.g. ligand and receptor branches) is proposed for cross-protein structural modeling. Specifically, in this module, Wasserstein affinity across graphs is calculated through cross-graph query (i.e. ligand (query) - receptor (key) or the converse), based on which the multi-head attention is derived to adaptively mine fine-grained cues of PPI sites. By stacking CGT modules, the two branches in CGT are co-evolved in a deep architecture during forward inference, hence being powerful and advantageous in cross-protein structural representation and fine-grained learning. We verify the effectiveness of our CPWT framework by conducting comprehensive experiments on multiple PPI datasets, and further visualize the learned fine-grained saliencies for intuitive understanding.
****

What Do Self-Supervised Vision Transformers Learn?
Namuk Park,Wonjae Kim,Byeongho Heo,Taekyung Kim,Sangdoo Yun
We present a comparative study on how and why contrastive learning (CL) and masked image modeling (MIM) differ in their representations and in their performance of downstream tasks. In particular, we demonstrate that self-supervised Vision Transformers (ViTs) have the following properties: (1) CL trains self-attentions to capture longer-range global patterns than MIM, such as the shape of an object, especially in the later layers of the ViT architecture. This CL property helps ViTs linearly separate images in their representation spaces. However, it also makes the self-attentions collapse into homogeneity for all query tokens and heads. Such homogeneity of self-attention reduces the diversity of representations, worsening scalability and dense prediction performance. (2) CL utilizes the low-frequency signals of the representations, but MIM utilizes high-frequencies. Since low- and high-frequency information respectively represent shapes and textures, CL is more shape-oriented and MIM more texture-oriented. (3) CL plays a crucial role in the later layers, while MIM mainly focuses on the early layers. Upon these analyses, we find that CL and MIM can complement each other and observe that even the simplest harmonization can help leverage the advantages of both methods. The code is available at https://github.com/naver-ai/cl-vs-mim.

****

Confident Sinkhorn Allocation for Pseudo-Labeling
Vu Nguyen,Sachin Farfade,Anton van den Hengel
Semi-supervised learning is a critical tool in reducing machine learning’s dependence on labeled data. It has been successfully applied to structure data, such as image and language data, by exploiting the inherent spatial and semantic structure therein with pretrained models or data augmentation. Some of these methods are no longer applicable for the data where domain structures are not available because the pretrained models or data augmentation can not be used.

Due to simplicity, existing pseudo-labeling (PL) methods can be widely used with out any domain assumption, but are vulnerable to noise samples and to greedy assignments given a predefined threshold which is typically unknown. This paper addresses this problem by proposing a Confident Sinkhorn Allocation (CSA), which assigns labels to only samples with high confidence scores and learns the best label allocation via optimal transport. CSA outperforms the current state-of-the-art in this practically important area of semi-supervised learning.

****

Adversarial Robustness based on Randomized Smoothing in Quantum Machine Learning

Aditya Sahdev,Mehul Kumar

We present an end-to-end Quantum Machine Learning algorithm that encodes a classical input into a Quantum Computing state and provides certified radius for a base classifier, with robustness guarantees based on randomized smoothing - current state-of-the-art defense against adversarial attacks. Classically, the number of samples, also the number of queries to the classifier, scale with $O(1/\epsilon^2)$ where ϵ is the desired error bound in expected value of the probability measure ρ defined over the randomized smoothing neighborhood. Our algorithm is designed to solve the same problem for a Quantum Computing classifier. We prove that number of queries to the classifier scale as $O(1/\epsilon)$ for the same confidence and error bound. We also present the unitary circuit corresponding to the quantum randomized smoothing algorithm, as well as the state preparation methods and circuits for smoothing distributions used to defend against common adversaries - modeled using l_0, l_1, l_2 norms, and other metrics. The results of comparison between the classical and simulation of the quantum algorithm are also discussed.

****

Multi-Vector Retrieval as Sparse Alignment

Yujie Qian,Jinhyuk Lee,Karthik Duddu,Zhuyun Dai,Tao Lei,Siddhartha Brahma,Iftekhar Naim,Vincent Y Zhao

Multi-vector retrieval models improve over single-vector dual encoders on many information retrieval tasks. In this paper, we cast the multi-vector retrieval problem as sparse alignment between query and document tokens. We propose ALIGNER, a novel multi-vector retrieval model that learns sparsified pairwise alignments between query and document tokens (e.g. `dog' vs. `puppy') and per-token unary saliences reflecting their relative importance for retrieval. We show that controlling the sparsity of pairwise token alignments often brings significant performance gains. While most factoid questions focusing on a specific part of a document require a smaller number of alignments, others requiring a broader understanding of a document favor a larger number of alignments. Unary saliences, on the other hand, decide whether a token ever needs to be aligned with others for retrieval (e.g. `kind' from `what kind of currency is used in new zealand'). With sparsified unary saliences, we are able to prune a large number of query and document token vectors and improve the efficiency of multi-vector retrieval. We learn the sparse unary saliences with entropy-regularized linear programming, which outperforms other methods to achieve sparsity. In a zero-shot setting, ALIGNER scores 51.1 nDCG@10, achieving a new retriever-only state-of-the-art on 13 tasks in the BEIR benchmark. In addition, adapting pairwise alignments with a few examples (<= 8) further improves the performance up to 15.7 points nDCG@10 for argument retrieval tasks. The unary saliences of ALIGNER helps us to keep only 20% of the document token representations with minimal performance loss. We further show that our model often produces interpretable alignments and significantly improves its performance when initialized from larger language models.

****

Sampled Transformer for Point Sets

Shidi Li,Christian Walder,Alexander Soen,Lexing Xie,miaomiao Liu

The sparse transformer can reduce the computational complexity of the self-attention layers to $O(n)$, whilst still being a universal approximator of continuous sequence-to-sequence functions. However, this permutation variant operation is not appropriate for direct application to sets. In this paper, we proposed an $O(n)$ complexity sampled transformer that can process point set elements directly without any additional inductive bias. Our sampled transformer introduces random element sampling, which randomly splits point sets into subsets, followed by applying a shared Hamiltonian self-attention mechanism to each subset. The overall attention mechanism can be viewed as a Hamiltonian cycle in the complete attention graph, and the permutation of point set elements is equivalent to randomly sampling Hamiltonian cycles. This mechanism implements a Monte Carlo simulation of the $O(n^2)$ dense attention connections. We show that it is a universal approximator for continuous set-to-set functions. Experimental results for classification and few-shot learning on point-clouds show comparable or better accuracy with significantly reduced computational complexity compared to the dense transformer or alternative sparse attention schemes.

****

Population-size-Aware Policy Optimization for Mean-Field Games

Pengdeng Li,Xinrun Wang,Shuxin Li,Hau Chan,Bo An

In this work, we attempt to bridge the two fields of finite-agent and infinite-agent games, by studying how the optimal policies of agents evolve with the number of agents (population size) in mean-field games, an agent-centric perspective in contrast to the existing works focusing typically on the convergence of the empirical distribution of the population. To this end, the premise is to obtain the optimal policies of a set of finite-agent games with different population sizes. However, either deriving the closed-form solution for each game is theoretically intractable, training a distinct policy for each game is computationally intensive, or directly applying the policy trained in a game to other games is sub-optimal. We address these challenges through the \textbf{P}opulation-size-\textbf{A}ware \textbf{P}olicy \textbf{O}ptimization (PAPO). Our contributions are three-fold. First, to efficiently generate efficient policies for games with different population sizes, we propose PAPO, which unifies two natural options (augmentation and hypernetwork) and achieves significantly better performance. PAPO consists of three components: i) the population-size encoding which transforms the original value of population size to an equivalent encoding to avoid training collapse, ii) a hypernetwork to generate a distinct policy for each game conditioned on the population size, and iii) the population size as an additional input to the generated policy. Next, we construct a multi-task-based training procedure to efficiently train the neural networks of PAPO by sampling data from multiple games with different population sizes. Finally, extensive experiments on multiple environments show the significant superiority of PAPO over baselines, and the analysis of the evolution of the generated policies further deepens our understanding of the two fields of finite-agent and infinite-agent games.

****

PartAfford: Part-level Affordance Discovery

Chao Xu,Yixin Chen,He Wang,Song-Chun Zhu,Yixin Zhu,Siyuan Huang

Understanding what objects could furnish for humans—learning object affordance—is the crux of bridging perception and action. In the vision community, prior work has primarily focused on learning object affordance with dense (e.g., at a per-pixel level) supervision. In stark contrast, we humans learn the object affordance without dense labels. As such, the fundamental question to devise a computational model is: What is the natural way to learn the object affordance from geometry with humanlike sparse supervision? In this work, we present the new task of part-level affordance discovery (PartAfford): Given only the affordance labels for each object, the machine is tasked to (i) decompose 3D shapes into parts and (ii) discover how each part of the object corresponds to a certain affordance category. We propose a novel learning framework that discovers part-level representations by leveraging only the affordance set supervision and geometric primitive regularization without dense supervision. To learn and evaluate PartAfford, we construct a part-level, cross-category 3D object affordance dataset, annotated with 24 affordance categories shared among >25, 000 objects. We demonstrate through extensive experiments that our method enables both the abstraction of 3D objects and part-level affordance discovery, with generalizability to difficult and cross-category examples. Further ablations reveal the contribution of each component.

****

On The Relative Error of Random Fourier Features for Preserving Kernel Distance

Kuan Cheng,Shaofeng H.-C. Jiang,Luojian Wei,Zhide Wei

The method of random Fourier features (RFF), proposed in a seminal paper by Rahimi and Recht (NIPS'07), is a powerful technique to find approximate low-dimensional representations of points in (high-dimensional) kernel space, for shift-invariant kernels. While RFF has been analyzed under various notions of error guarantee, the ability to preserve the kernel distance with \emph{relative} error is less understood. We show that for a significant range of kernels, including the well-known Laplacian kernels, RFF cannot approximate the kernel distance with small relative error using low dimensions. We complement this by showing as long as the shift-invariant kernel is analytic, RFF with $\mathrm{poly}(\epsilon^{-1} \log n)$ dimensions achieves ϵ-relative error for pairwise kernel distance of n points, and the dimension bound is improved to $\mathrm{poly}(\epsilon^{-1}\log k)$ for the specific application of kernel k-means. Finally, going beyond RFF, we make the first step towards data-oblivious dimension-reduction for general shift-invariant kernels, and we obtain a similar $\mathrm{poly}(\epsilon^{-1} \log n)$ dimension bound for Laplacian kernels. We also validate the dimension-error tradeoff of our methods on simulated datasets, and they demonstrate superior performance compared with other popular methods including random-projection and Nystr\"{o}m methods.

****

UTC-IE: A Unified Token-pair Classification Architecture for Information Extraction

Hang Yan,Yu Sun,Xiaonan Li,Yunhua Zhou,Xuanjing Huang,Xipeng Qiu

Information Extraction (IE) spans several tasks with different output structures, such as named entity recognition, relation extraction and event extraction. Previously, those tasks were solved with different models because of diverse task output structures. Through re-examining IE tasks, we find that all of them can be interpreted as extracting spans and span relations. We propose using the start and end token of a span to pinpoint the span in texts, and using the start-to-start and end-to-end token pairs of two spans to determine the relation. Hence, we can unify all IE tasks under the same token-pair classification formulation. Based on the reformulation, we propose a \textbf{U}nified \textbf{T}oken-pair \textbf{C}lassification architecture for \textbf{I}nformation \textbf{E}xtraction (\textbf{UTC-IE}), where we introduce Plusformer on top of the token-pair feature matrix. Specifically, it models axis-aware interaction with plus-shaped self-attention and local interaction with Convolutional Neural Network over token pairs. Experiments show that our approach outperforms task-specific and unified models on all tasks in 10 datasets, and achieves better or comparable results on 2 joint IE datasets. Moreover, UTC-IE speeds up over state-of-the-art models on IE tasks significantly in most datasets, which verifies the effectiveness of our architecture.

****

Linear Convergence of Decentralized FedAvg for Non-Convex Objectives: The Interpolation Regime

Shruti P Maralappanavar,Prashant Khanduri,Bharath B N

In the age of Bigdata, Federated Learning (FL) provides machine learning (ML) practitioners with an indispensable tool for solving large-scale learning problems. FL is a distributed optimization paradigm where multiple nodes each having access to a local dataset collaborate (with or without a server) to solve a joint problem. Federated Averaging (FedAvg) although the algorithm of choice for many FL applications is not very well understood especially in the interpolation regime, a common phenomenon observed in modern overparameterized neural networks. In this work, we address this challenge and perform a thorough theoretical performance analysis of FedAvg in the interpolation regime for training of overparameterized neural networks. Specifically, we analyze the performance of FedAvg in two settings: (i) {\em[Server]}: When the network has access to a server that coordinates the information sharing among nodes, and (ii) {\em[Decentralized]:} The serverless setting, where the local nodes communicate over an undirected graph. We consider a class of non-convex functions satisfying the Polyak-Lojasiewicz (PL) condition, a condition that is satisfied by overparameterized neural networks. For the first time, we establish that FedAvg under both {\em Server} and {\em Decentralized} settings achieve linear convergence rates of $\mathcal{O}(T^{3/2} \log (1/{\epsilon}))$ and $\mathcal{O}({T^2} \log ({1}/{\epsilon}))$, respectively, where ϵ is the desired solution accuracy, and T is the number of local updates at each node. In contrast to the standard FedAvg analysis, our work does not require bounded heterogeneity, variance, and gradient assumptions. Instead, we show that sample-wise (and local) smoothness of the local loss functions suffice to capture the effect of heterogeneity in FL training. We use a novel application of induction to prove the linear convergence in the {\em Decentralized} setting, which can be of independent interest. Finally, we conduct experiments on multiple real datasets to corroborate our theoretical findings.

****

Rethinking Missing Modality Learning: From a Decoding View
Tao Jin,Zhou Zhao

Conventional pipeline of multimodal learning consists of three stages, including encoding, fusion, and decoding. Most existing methods under missing modality condition focus on the first stage and aim to learn the modality invariant representation or reconstruct missing features. However, these methods rely on strong assumptions (i.e., all the pre-defined modalities are available for each input sample during training and the number of modalities is fixed). To solve this problem, we propose a simple yet effective method called Interaction Augmented Prototype Decomposition (IPD) for a more general setting, where the number of modalities is arbitrary and there are various incomplete modality conditions happening in both training and inference phases, even there are unseen testing conditions. Different from the previous methods, we improve the decoding stage. Concretely, IPD jointly learns the common and modality-specific task prototypes. Considering that the number of missing modality conditions scales exponentially with the number of modalities ${\bf O}({\text 2^n})$ and different conditions may have implicit interaction, the low-rank partial prototype decomposition with enough theoretical analysis is employed for modality-specific components to reduce the complexity. The decomposition also can promote unseen generalization with the modality factors of existing conditions. To simulate the low-rank setup, we further constrain the explicit interaction of specific modality conditions by employing disentangled contrastive constraints. Extensive results on the newly-created benchmarks of multiple tasks illustrate the effectiveness of our proposed model.

****

UNDERSTANDING PURE CLIP GUIDANCE FOR VOXEL GRID NERF MODELS
Han-Hung Lee,Angel X Chang

We explore the task of text to 3D object generation using CLIP. Specifically, we use CLIP for guidance without access to any datasets, a setting we refer to as pure CLIP guidance. While prior work has adopted this setting, there is no systematic study of mechanics for preventing adversarial generations within CLIP. We illustrate how different image-based augmentations prevent the adversarial generation problem, and how the generated results are impacted. We test different CLIP model architectures and show that ensembling different models for guidance can prevent adversarial generations within bigger models and generate sharper results. Furthermore, we implement an implicit voxel grid model to show how neural networks provide an additional layer of regularization, resulting in better geometrical structure and coherency of generated objects. Compared to prior work, we achieve more coherent results with higher memory efficiency and faster training speeds.

****

Task-Agnostic Online Meta-Learning in Non-stationary Environments
Daouda Sow,Sen Lin,Yingbin Liang,Junshan Zhang

Online meta-learning has recently emerged as a marriage between batch meta-learning and online learning, for achieving the capability of quick adaptation on new tasks in a lifelong manner. However, most existing approaches focus on the restrictive setting where the distribution of the online tasks remains fixed with kn own task boundaries. In this work we relax these assumptions and propose a novel algorithm for task-agnostic online meta-learning in non-stationary environments. More specifically, we first propose two simple but effective detection mechanisms of task switches and distribution shift based on empirical observations, which serve as a key building block for more elegant online model updates in our algorithm: the task switch detection mechanism allows reusing of the best model available for the current task at hand, and the distribution shift detection mechanism differentiates the meta model update so as to preserve the knowledge for in-distribution tasks and quickly learn the new knowledge for out-of-distribution tasks.

Motivated by the recent advance in online learning, our online meta model updates are based only on the current data, which eliminates the need of storing previous data as required in most existing methods. This crucial choice is also well supported by our theoretical analysis of dynamic regret in online meta-learning, where a sublinear regret can be achieved by updating the meta model at each round using the current data only. Empirical studies on three different benchmarks clearly demonstrate the significant advantage of our algorithm over related baseline approaches.

****

Meta-Weighted Language Model Tuning for Augmentation-Enhanced Few-Shot Learning

Yu Meng,Martin Michalski,Jiaxin Huang,Yu Zhang,Tarek Abdelzaher,Jiawei Han

Recent studies have revealed the intriguing few-shot learning ability of pretrained language models (PLMs): They can quickly adapt to a new task when fine-tuned on a small amount of labeled data formulated as prompts, without requiring abundant task-specific annotations. Despite their promising performance, most existing few-shot approaches that only learn from the small training set still underperform fully supervised training by nontrivial margins. In this work, we study few-shot learning with PLMs from a different perspective: We first tune an autoregressive PLM on the few-shot samples and then use it as a generator to synthesize a large amount of novel training samples which augment the original training set. To encourage the generator to produce label-discriminative samples, we train it via weighted maximum likelihood where the weight of each token is automatically adjusted based on a discriminative meta-learning objective. A classification PLM can then be fine-tuned on both the few-shot and the synthetic samples with regularization for better generalization and stability. Our approach FewGen achieves an overall better result across seven classification tasks of the GLUE benchmark than existing few-shot learning methods.

****

Online Reinforcement Learning via Posterior Sampling of Policy

Shuqing Shi,zhiqiang xu,Fan Zhang,Zhiyou Yang,Yali Du,Hong Qu

We propose a Reward-Weighted Posterior Sampling of Policy (RWPSP) algorithm to tackle the classic trade-off problem between exploration and exploitation under finite Markov decision processes (MDPs). The Thompson sampling method so far has only considered posterior sampling over transition probabilities, which is hard to gain the globally sub-optimal rewards. RWPSP runs posterior sampling over stationary policy distributions instead of transition probabilities, and meanwhile keeps transition probabilities updated. Particularly, we leverage both relevant count functions and reward-weighting to online update the policy posterior, aiming to balance between local and long-term policy distributions for a globally near-optimal game value. Theoretically, we establish a bound of $\tilde{\mathcal{O}}(\Gamma\sqrt{T}/S^{2})$\footnote{The symbol $\tilde{\mathcal{O}}$ hides logarithmic factors.} on the total regret in time horizon T with $\Gamma/S^{2} < D\sqrt{SA}$ satisfied in general, where S and A represents the sizes of state and action spaces, respectively, D the diameter. This matches the best regret bound thus far for MDPs. Experimental results corroborate our theoretical results and show the advantage of our algorithm over the state of the art in terms of efficiency.

****

DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing

Pengcheng He,Jianfeng Gao,Weizhu Chen

This paper presents a new pre-trained language model, NewModel, which improves the original DeBERTa model by replacing mask language modeling (MLM) with replaced token detection (RTD), a more sample-efficient pre-training task. Our analysis shows that vanilla embedding sharing in ELECTRA hurts training efficiency and model performance. This is because the training losses of the discriminator and the generator pull token embeddings in different directions, creating the “tug-of-war” dynamics. We thus propose a new gradient-disentangled embedding sharing method that avoids the tug-of-war dynamics, improving both training efficiency and the quality of the pre-trained model. We have pre-trained NewModel using the same settings as DeBERTa to demonstrate its exceptional performance on a wide range of downstream natural language understanding (NLU) tasks. Taking the GLUE benchmark with eight tasks as an example, the NewModel Large model achieves a 91.37% average score, which is 1.37% over DeBERTa and 1.91% over ELECTRA, setting a new state-of-the-art (SOTA) among the models with a similar structure. Furthermore, we have pre-trained a multi-lingual model mNew-Model and observed a larger improvement over strong baselines compared to English models. For example, the mNewModel Base achieves a 79.8% zero-shot cross-lingual accuracy on XNLI and a 3.6% improvement over XLM-R Base, creating a new SOTA on this benchmark. We will make our model and code publicly available.

****

Weakly Supervised Neuro-Symbolic Image Manipulation via Multi-Hop Complex Instructions

Harman Singh,Poorva Garg,Mohit Gupta,Kevin Shah,Arnab Kumar Mondal,Dinesh Khandelwal,Parag Singla,Dinesh Garg

We are interested in image manipulation via natural language text – a task that is extremely useful for multiple AI applications but requires complex reasoning over multi-modal spaces. Recent work on neuro-symbolic approaches (Mao et al., 2019) (NSCL) has been quite effective for solving VQA as they offer better modularity, interpretability, and generalizability. We extend NSCL for the image manipulation task and propose a solution referred to as NeuroSIM. Previous work either requires supervised training data in the form of manipulated images or can only deal with very simple reasoning instructions over single object scenes. In contrast, NeuroSIM can perform complex multi-hop reasoning over multi-object scenes and only requires weak supervision in the form of annotated data for VQA. NeuroSIM parses an instruction into a symbolic program, based on a Domain Specific Language (DSL) comprising of object attributes and manipulation operations, that guides the manipulation. We design neural modules for manipulation, as well as novel loss functions that are capable of testing the correctness of manipulated object and scene graph representations via query networks trained merely on VQA data. An image decoder is trained to render the final image from the manipulated scene graph. Extensive experiments demonstrate that NeuroSIM, without using target images as supervision, is highly competitive with SOTA baselines that make use of supervised data for manipulation.

****

Graph Neural Networks for Aerodynamic Flow Reconstruction from Sparse Sensing

Gregory Duthé,Imad Abdallah,Sarah Barber,Eleni Chatzi

Sensing the fluid flow around an arbitrary geometry entails extrapolating from the physical quantities perceived at its surface in order to reconstruct the features of the surrounding fluid. This is a challenging inverse problem, yet one that if solved could have a significant impact on many engineering applications. The exploitation of such an inverse logic has gained interest in recent years with the advent of widely available cheap but capable MEMS-based sensors. When combined with novel data-driven methods, these sensors may allow for flow reconstruction around immersed structures, benefiting applications such as unmanned airborne/underwater vehicle path planning or control and structural health monitoring of wind turbine blades. In this work, we train deep reversible Graph Neural Networks (GNNs) to perform flow sensing (flow reconstruction) around two-dimensional aerodynamic shapes: airfoils. Motivated by recent work, which has shown that GNNs can be powerful alternatives to mesh-based forward physics simulators, we i mplement a Message-Passing Neural Network to simultaneously reconstruct both the pressure and velocity fields surrounding simulated airfoils based on their surface pressure distributions, whilst additionally gathering useful farfield properties in the form of context vectors. We generate a unique dataset of Computational Fluid Dynamics simulations by simulating random, yet meaningful combinations of input boundary conditions and airfoil shapes. We show that despite the challenges associated with reconstructing the flow around arbitrary airfoil geometries in high Reynolds turbulent inflow conditions, our framework is able to generalize well to unseen cases.

****

Learning Binary Networks on Long-Tailed Distributions

Jihun Kim,Dahyun Kim,Hyung Rok Jung,TaeIL Oh,Jonghyun Choi

In deploying deep models to real world scenarios, there are a number of issues including computational resource constraints and long-tailed data distributions. For the first time in the literature, we address the combined challenge of learning long-tailed distributions under the extreme resource constraints of using binary networks as backbones. Specifically, we propose a framework of calibrating off-the-shelf pretrained full precision weights that are learned on $\textit{non-long-tailed}$ distributions when training binary networks on long-tailed datasets. In the framework, we additionally propose a novel adversarial balancing and a multi-resolution learning method for better generalization to diverse semantic domains and input resolutions. We conduct extensive empirical evaluations on 15 datasets including newly derived long-tailed datasets from existing balanced datasets, which is the largest benchmark in the literature. Our empirical studies show that our proposed method outperforms prior arts by large margins, $\textit{e.g.}$, at least $+14.33\%$ on average.

****

Backdoor Mitigation by Correcting Activation Distribution Alteration

Xi Li,Zhen Xiang,George Kesidis,Bo Li,David J. Miller

Backdoor (Trojan) attacks are an important type of adversarial exploit against deep neural networks (DNNs), wherein a test instance is (mis)classified to the attacker's target class whenever a backdoor trigger is present. In this paper, we reveal and analyze an important property of backdoor attacks: a successful attack causes an alteration in the distribution of internal layer activations for backdoor-trigger instances, compared to that for clean instances. Even more importantly, we find that instances with the backdoor trigger will be correctly classified to their original source classes if this distribution alteration is reversed. Based on our observations, we propose an efficient and effective method that achieves post-training backdoor mitigation by correcting the distribution alteration using reverse-engineered triggers. Notably, our method does not change any trainable parameters of the DNN, but achieves generally better mitigation performance than existing methods that do require intensive DNN parameter tuning. It also efficiently detects test instances with the trigger, which may help to catch adversarial entities.

****

Local Distance Preserving Auto-encoders using Continuous k-Nearest Neighbours Graphs

Nutan Chen,Patrick van der Smagt,Botond Cseke

Auto-encoder models that preserve similarities in the data are a popular tool in representation learning. In this paper we introduce several auto-encoder models that preserve local distances when mapping from the data space to the latent space. We use a local distance-preserving loss that is based on the continuous k-nearest neighbours graph which is known to capture topological features at all scales simultaneously. To improve training performance, we formulate learning as a constraint optimisation problem with local distance preservation as the main objective and reconstruction accuracy as a constraint. We generalise this approach to hierarchical variational auto-encoders thus learning generative models with geometrically consistent latent and data spaces. Our method provides state-of-the-art or comparable performance across several standard datasets and evaluation metrics.

****

Clustering for directed graphs using parametrized random walk diffusion kernels

Harry Sevi,Matthieu Jonckheere,Argyris Kalogeratos

Clustering based on the random walk operator has been proven effective for undirected graphs, but its generalization to directed graphs (digraphs) is much more challenging. Although the random walk operator is well-defined for digraphs, in most cases such graphs are not strongly connected, and hence the associated random walks are not irreducible, which is a crucial property for clustering that exists naturally in the undirected setting. To remedy this, the usual workaround is to either naively symmetrize the adjacency matrix or to replace the natural random walk operator by the teleporting random walk operator, but this can lead to the loss of valuable information carried by edge directionality. In this paper, we introduce a new clustering framework, the Parametrized Random Walk Diffusion Kernel Clustering (P-RWDKC), which is suitable for handling both directed and undirected graphs. Our framework is based on the diffusion geometry and the generalized spectral clustering framework. Accordingly, we propose an algorithm that automatically reveals the cluster structure at a given scale, by considering the random walk dynamics associated with a parametrized kernel operator, and by estimating its critical diffusion time. Experiments on K-NN graphs constructed from real-world datasets and real-world graphs show that our clustering approach performs well in all tested cases, and outperforms existing approaches in most of them.

****

Poisoning Generative Models to Promote Catastrophic Forgetting

Siteng Kang,Zhan Shi,Xinhua Zhang

Generative models have grown into the workhorse of many state-of-the-art machine learning methods. However, their vulnerability under poisoning attacks has been largely understudied. In this work, we investigate this issue in the context of continual learning, where generative replayers are utilized to tackle catastrophic forgetting. By developing a novel customization of dirty-label input-aware backdoor to the online setting, our attacker manages to stealthily promote forgetting while retaining high accuracy at the current task and sustaining strong defenders. Our approach taps into an intriguing property of generative models, namely that they cannot well capture input-dependent triggers. Experiments on four standard datasets corroborate the poisoner’s effectiveness.

****

Squeeze Training for Adversarial Robustness

Qizhang Li,Yiwen Guo,Wangmeng Zuo,Hao Chen

The vulnerability of deep neural networks (DNNs) to adversarial examples has attracted great attention in the machine learning community. The problem is related to non-flatness and non-smoothness of normally obtained loss landscapes. Training augmented with adversarial examples (a.k.a., adversarial training) is considered as an effective remedy. In this paper, we highlight that some collaborative examples, nearly perceptually indistinguishable from both adversarial and benign examples yet show extremely lower prediction loss, can be utilized to enhance adversarial training. A novel method is therefore proposed to achieve new state-of-the-arts in adversarial robustness. Code: https://github.com/qizhangli/ST-AT.

****

Knowledge Unlearning for Mitigating Privacy Risks in Language Models

Joel Jang,Dongkeun Yoon,Sohee Yang,Sungmin Cha,Moontae Lee,Lajanugen Logeswaran,
Minjoon Seo

Pretrained Language Models (LMs) memorize a vast amount of knowledge during initial pretraining, including information that may violate the privacy of personal lives and identities. Previous work addressing privacy issues for language models has mostly focused on data preprocessing and differential privacy methods, both requiring re-training the underlying LM. We propose knowledge unlearning as an alternative method to reduce privacy risks for LMs post hoc. We show that simply applying the unlikelihood training objective to target token sequences is effective at forgetting them with little to no degradation of general language model ing performances; it sometimes even substantially improves the underlying LM with just a few iterations. We also find that sequential unlearning is better than trying to unlearn all the data at once and that unlearning is highly dependent on which kind of data (domain) is forgotten. By showing comparisons with a previous data preprocessing method known to mitigate privacy risks for LMs, we show that unlearning can give a stronger empirical privacy guarantee in scenarios where the data vulnerable to extraction attacks are known a priori while being orders of magnitude more computationally efficient. We release the code and dataset needed to replicate our results at http://www.omitted.link/.

****

Revisiting the Activation Function for Federated Image Classification
Jaewoo Shin,Taehyeon Kim,Se-Young Yun

Federated learning (FL) has become one of the most popular distributed machine learning paradigms; these paradigms enable training on a large corpus of decentralized data that resides on devices. The recent evolution in FL research is mainly credited to the refinements in training procedures by developing the optimization methods. However, there has been little verification of other technical improvements, especially improvements to the activation functions (e.g., ReLU), that are widely used in the conventional centralized approach (i.e., standard data-centric optimization). In this work, we verify the effectiveness of activation functions in various federated settings. We empirically observe that off-the-shelf activation functions that are used in centralized settings exhibit a totally different performance trend than do federated settings. The experimental results demonstrate that HardTanh achieves the best accuracy when severe data heterogeneity or low participation rate is present. We provide a thorough analysis to investigate why the representation powers of activation functions are changed in a federated setting by measuring the similarities in terms of weight parameters and representations. Lastly, we deliver guidelines for selecting activation functions in both a cross-silo setting (i.e., a number of clients <= 20) and a cross-device setting (i.e., a number of clients >= 100). We believe that our work provides benchmark data and intriguing insights for designing models FL models.

****

Open-domain Visual Entity Linking
Hexiang Hu,Yi Luan,Urvashi Khandelwal,Mandar Joshi,Kenton Lee,Kristina Toutanova,Ming-Wei Chang

We introduce the task of Open-domain Visual Entity Linking (OVEN), targeting a wide range of entities including animals, plants, buildings, locations and much more. Given an image (e.g., an image of an aircraft), a text query (`What is the model?' or `What is the airline?'), and a multi-modal knowledge base (e.g., Wikipedia), the goal is to link to an entity (Boeing-777 or EVA Air) out of all entities in the knowledge base. We build a benchmark dataset (OVEN-wiki), by repurposing 14 existing image classification, image retrieval, and visual QA datasets. We link all existing labels to Wikipedia entities when possible, using a state-of-the-art entity linking system and human annotators, creating a diverse and unified label space. OVEN is a rich and challenging task, which requires models to recognize and link visual content to both a small set of seen entities as well as a much larger set of unseen entities (e.g., unseen aircraft models). OVEN also requires models to generalize to previously unseen intents that may require more fine-grained reasoning (`Who manufactured the aircraft in the back?'). We build strong baselines based on state-of-the-art pre-trained models and find that current pre-trained models struggle to address the challenges posed by OVEN. We hope OVEN will inspire next-generation pre-training techniques and pave the way to future knowledge-intensive vision tasks.

****

Robustify Transformers with Robust Kernel Density Estimation
Xing Han,Tongzheng Ren,Tan Minh Nguyen,Khai Nguyen,Joydeep Ghosh,Nhat Ho

Recent advances in Transformer architecture have empowered its empirical success in various tasks across different domains. However, existing works mainly focus on improving the standard accuracy and computational cost, without considering the robustness of contaminated samples. Existing work (Nguyen et al, 2022, Fouri erFormer) has shown that the self-attention mechanism, which is the center of the Transformer architecture, can be viewed as a non-parametric estimator based on the well-known kernel density estimation (KDE). This motivates us to leverage the robust kernel density estimation (RKDE) in the self-attention mechanism, to alleviate the issue of the contamination of data by down-weighting the weight of bad samples in the estimation process. The modified self-attention mechanism can be incorporated into different Transformer variants. Empirical results on language modeling and image classification tasks demonstrate the effectiveness of this approach.

****

Pushing the Accuracy-Group Robustness Frontier with Introspective Self-play

Jeremiah Zhe Liu,Krishnamurthy Dj Dvijotham,Jihyeon Lee,Quan Yuan,Balaji Lakshminarayanan,Deepak Ramachandran

Standard empirical risk minimization (ERM) training can produce deep neural network (DNN) models that are accurate on average but under-perform in under-represented population subgroups, especially when there are imbalanced group distributions in the long-tailed training data. Therefore, approaches that improve the accuracy - group robustness trade-off frontier of a DNN model (i.e. improving worst-group accuracy without sacrificing average accuracy, or vice versa) is of crucial importance. Uncertainty-based active learning (AL) can potentially improve the frontier by preferentially sampling underrepresented subgroups to create a more balanced training dataset. However, the quality of uncertainty estimates from modern DNNs tend to degrade in the presence of spurious correlations and dataset bias, compromising the effectiveness of AL for sampling tail groups. In this work, we propose Introspective Self-play (ISP), a simple approach to improve the uncertainty estimation of a deep neural network under dataset bias, by adding an auxiliary introspection task requiring a model to predict the bias for each data point in addition to the label. We show that ISP provably improves the bias-awareness of the model representation and the resulting uncertainty estimates. On two real-world tabular and language tasks,ISP serves as a simple “plug-in” for AL model training, consistently improving both the tail-group sampling rate and the final accuracy-fairness trade-off frontier of popular AL methods.

****

Learning to Predict Parameter for Unseen Data

Shiye Wang,Kaituo Feng,Changsheng Li,Ye Yuan,Guoren Wang

Typical deep learning models depend heavily on large amounts of training data and resort to an iterative optimization algorithm (e.g., SGD or Adam) for learning network parameters, which makes the training process very time- and resource-intensive. In this paper, we propose a new training paradigm and formulate network parameter training into a prediction task: given a network architecture, we observe there exists correlations between datasets and their corresponding optimal network parameters, and explore if we can learn a hyper-mapping between them to capture the relations, such that we can directly predict the parameters of the network for a new dataset never seen during the training phase. To do this, we put forward a new hypernetwork with the purpose of building a mapping between datasets and their corresponding network parameters, and then predict parameters for unseen data with only a single forward propagation of the hypernetwork. At its heart, our model benefits from a series of GRU sharing weights to capture the dependencies of parameters among different layers in the network. Extensive experimental studies are performed and experimental results validate our proposed method achieves surprisingly good efficacy. For instance, it takes 119 GPU seconds to train ResNet-18 using Adam from scratch and obtain a top-1 accuracy of 74.56%, while our method costs only 0.5 GPU seconds to predict the network parameters of ResNet-18 achieving comparable performance (73.33%), more than 200 times faster than the traditional training paradigm.

****

UNREAL: Unlabeled Nodes Retrieval and Labeling for Heavily-imbalanced Node Classification

Liang Yan,Shengzhong Zhang,Bisheng Li,min zhou,Zengfeng Huang

Extremely skewed label distributions are common in real-world node classificatio

n tasks. If not dealt with appropriately, it significantly hurts the performance of GNNs on minority classes. Due to the practical importance, there have been a series of recent researches devoted to this challenge. Existing over-sampling techniques smooth the label distribution by generating ''fake'' minority nodes and synthesize their features and local topology, which largely ignore the rich information of unlabeled nodes on graphs. Recent methods based on loss function modification re-weight different samples or change classification margins, which achieve good performance. However, representative methods need label information to estimate the distance of each node to its class center, which is unavailable on unlabeled nodes. In this paper, we propose UNREAL, which is an iterative over-sampling method. The first key difference is that we only add unlabeled nodes instead of synthetic nodes, which eliminates the challenge of feature and neighborhood generation. To select which unlabeled nodes to add, we propose geometric ranking, which ranks unlabeled nodes based on unsupervised learning results in the node embedding space. Finally, we identify the issue of geometric imbalance in the embedding space and provide a simple metric to filter out geometrically imbalanced nodes. Extensive experiments on real-world benchmark datasets are conducted, and the empirical results show that our method significantly outperforms current state-of-the-art methods consistent on different datasets with different imbalance ratios.
****

On Nullspace of Vision Transformers and What Does it Tell Us?

Aditya Singh,Haohan Wang

Nullspace of a linear mapping is the subspace which is mapped to the zero vector. For a linear map, adding an element of the nullspace to its input has no effect on the output of the mapping. We position this work as an exposition towards answering one simple question, ``Does a vision transformer have a non-trivial nullspace?" If TRUE, this would imply that adding elements from this non-trivial nullspace to an input will have no effect on the output of the network. This finding can eventually lead us closer to understanding the generalization properties of vision transformers. In this paper, we first demonstrate that provably a non-trivial nullspace exists for a particular class of vision transformers. This proof is drawn by simply computing the nullspace of the patch embedding matrices. We extend this idea to the non-linear layers of the vision transformer and show that it is possible to learn a non-linear counterpart to the nullspace via simple optimisations for any vision transformer. Subsequently, we perform studies to understand robustness properties of ViTs under nullspace noise. Under robustness, we investigate prediction stability, and (network and interpretation) fooling properties of the noise. Lastly, we provide image watermarking as an application of nullspace noise.
****

Max-Margin Works while Large Margin Fails: Generalization without Uniform Convergence

Margalit Glasgow,Colin Wei,Mary Wootters,Tengyu Ma

A major challenge in modern machine learning is theoretically understanding the generalization properties of overparameterized models. Many existing tools rely on uniform convergence (UC), a property that, when it holds, guarantees that the test loss will be close to the training loss, uniformly over a class of candidate models. Nagarajan and Kolter (2019) show that in certain simple linear and neural-network settings, any uniform convergence bound will be vacuous, leaving open the question of how to prove generalization in settings where UC fails. Our main contribution is proving novel generalization bounds in two such settings, one linear, and one non-linear. We study the linear classification setting of Nagarajan and Kolter (2019), and a quadratic ground truth function learned via a two-layer neural network in the non-linear regime. We prove a new type of margin bound showing that above a certain signal-to-noise threshold, any near-max-margin classifier will achieve almost no test loss in these two settings. Our results show that near-max-margin is important: while any model that achieves at least a $(1 - \epsilon)$-fraction of the max-margin generalizes well, a classifier achieving half of the max-margin may fail terribly. Our analysis provides insight on

why memorization can coexist with generalization: we show that in this challenging regime where generalization occurs but UC fails, near-max-margin classifiers simultaneously contain some generalizable components and some overfitting components that memorize the data. The presence of the overfitting components is enough to preclude UC, but the near-extremal margin guarantees that sufficient generalizable components are present.
****

The batch size can affect inference results
Yunkyung Park,Kyungsoo Kim,Deok-kyu Jang

When performing matrix multiplication using GPUs, the cuBLAS library is commonly used for computational efficiency. Because of the cuBLAS’ heuristics, a vast, deep neural network model with GPUs may produce different test results owing to the batch sizes in both the training and inference stages. In this paper, we show that the batch size affects the inference results of deep neural network models. Our test models were the well-known bidirectional encoder representations from transformers (BERT) and generative pre-trained transformer (GPT) natural language processing (NLP) models, and the super-resolution generative adversarial network (SRGAN) image generation model in FP32 and TF32. In the TF32 setting, the evaluation loss in BERT using the general language understanding evaluation (GLUE) data sometimes varied for different batch sizes. The GPT generated sentences depending on batch size, and we show the logit's mean square error by increasing the token length. The SRGAN model produced different images from batch to batch. However, these phenomena were not observed under the FP32 setting. Therefore, the batch size must be carefully managed in large-sized deep neural networks under the TF32 setting.
****

Asymptotic Instance-Optimal Algorithms for Interactive Decision Making
Kefan Dong,Tengyu Ma

Past research on interactive decision making problems (bandits, reinforcement learning, etc.) mostly focuses on the minimax regret that measures the algorithm's performance on the hardest instance. However, an ideal algorithm should adapt to the complexity of a particular problem instance and incur smaller regrets on easy instances than worst-case instances. In this paper, we design the first asymptotic instance-optimal algorithm for general interactive decision making problems with finite number of decisions under mild conditions. On every instance f, our algorithm outperforms all consistent algorithms (those achieving non-trivial regrets on all instances), and has asymptotic regret $\mathcal{C}(f) \ln n$, where $\mathcal{C}(f)$ is an exact characterization of the complexity of f. The key step of the algorithm involves hypothesis testing with active data collection. It computes the most economical decisions with which the algorithm collects observations to test whether an estimated instance is indeed correct; thus, the complexity $\mathcal{C}(f)$ is the minimum cost to test the instance f against other instances. Our results, instantiated on concrete problems, recover the classical gap-dependent bounds for multi-armed bandits and prior works on linear bandits, and improve upon the previous best instance-dependent upper bound for reinforcement learning.
****

GRAPHSENSOR: A Graph Attention Network for Time-Series Sensor Data
jianchao lu,Yuzhe Tian,Quan Z. Sheng,Xi Zheng

Our work focuses on the exploration of the internal relationships of signals in an individual sensor. In particular, we address the problem of not being able to evaluate such inter-sensor relationships due to missing rich and explicit feature representation. To solve this problem, we propose GRAPHSENSOR, a graph attention network, with a shared-weight convolution feature encoder to generate the signal segments and learn the internal relationships between them. Furthermore, we enrich the representation of the features by utilizing a multi-head approach when creating the internal relationship graph. Compared with traditional multi-head approaches, we propose a more efficient convolution-based multi-head mechanism, which only requires 56% of model parameters compared with the best multi-head baseline as demonstrated in the experiments. Moreover, GRAPHSENSOR is capable of

achieving the state-of-the-art performance in the electroencephalography dataset and improving the accuracy by 13.8% compared to the best baseline in an inertial measurement unit (IMU) dataset.
****

ProsodyBERT: Self-Supervised Prosody Representation for Style-Controllable TTS
Yushi Hu,Chunlei Zhang,Jiatong Shi,Jiachen Lian,Mari Ostendorf,Dong Yu
We propose ProsodyBERT, a self-supervised approach to learning prosody representations from raw audio. Different from most previous works, which use information bottlenecks to disentangle prosody features from speech content and speaker information, we perform an offline clustering of speaker-normalized prosody-related features (energy, pitch, their dynamics, etc.) and use the cluster labels as targets for HuBERT-like masked unit prediction. A span boundary loss is also introduced to capture long-range prosodic information. We demonstrate the effectiveness of ProsodyBERT on a multi-speaker style-controllable text-to-speech (TTS) system. Experiments show that the TTS system trained with ProsodyBERT features can generate natural and expressive speech samples, surpassing the model supervised by energy and pitch on subjective human evaluation. Also, the style and expressiveness of synthesized audio can be controlled by manipulating the prosody features. In addition, We achieve new state-of-the-art results on the IEMOCAP emotion recognition task by combining our prosody features with HuBERT features, showing that ProsodyBERT is complementary to popular pretrained speech self-supervised models.
****

FedDebias: Reducing the Local Learning Bias Improves Federated Learning on Heterogeneous Data
Yongxin Guo,Xiaoying Tang,Tao Lin
Federated Learning (FL) is a machine learning paradigm that learns from data kept locally to safeguard the privacy of clients, whereas local SGD is typically employed on the clients' devices to improve communication efficiency. However, such a scheme is currently constrained by the slow and unstable convergence induced by clients' heterogeneous data.
In this work, we identify three under-explored phenomena of the biased local learning that may explain these challenges caused by local updates in supervised FL.
As a remedy, we propose FedDebias, a novel unified algorithm that reduces the local learning bias on features and classifiers to tackle these challenges.
FedDebias consists of two components:
The first component alleviates the bias in the local classifiers by balancing the output distribution of models.
The second component learns client invariant features that are close to global features but considerably distinct from those learned from other input distributions.
In a series of experiments, we show that FedDebias consistently outperforms other SOTA FL and domain generalization (DG) baselines, in which both two components have individual performance gains.
****

CRISP: Curriculum inducing Primitive Informed Subgoal Prediction for Hierarchical Reinforcement Learning
Utsav Singh,Vinay P Namboodiri
Hierarchical reinforcement learning is a promising approach that uses temporal abstraction to solve complex long horizon problems. However, simultaneously learning a hierarchy of policies is unstable as it is challenging to train higher-level policy when the lower-level primitive is non-stationary. In this paper, we propose to generate a curriculum of achievable subgoals for evolving lower-level primitives using reinforcement learning and imitation learning. The lower level primitive periodically performs data relabeling on a handful of expert demonstrations using our primitive informed parsing method. We derive expressions to bound the sub-optimality of our method and develop a practical algorithm for hierarchical reinforcement learning. Since our approach uses a handful of expert demonstrations, it is suitable for most real world robotic control tasks. Experimental

results on complex maze navigation and robotic manipulation environments show that inducing hierarchical curriculum learning significantly improves sample efficiency, and results in better learning of goal conditioned policies in complex temporally extended tasks.

****

Near-Optimal Deployment Efficiency in Reward-Free Reinforcement Learning with Linear Function Approximation

Dan Qiao,Yu-Xiang Wang

We study the problem of deployment efficient reinforcement learning (RL) with linear function approximation under the \emph{reward-free} exploration setting. This is a well-motivated problem because deploying new policies is costly in real-life RL applications. Under the linear MDP setting with feature dimension d and planning horizon H, we propose a new algorithm that collects at most $\widetilde{O}(\frac{d^2H^5}{\epsilon^2})$ trajectories within H deployments to identify ϵ-optimal policy for any (possibly data-dependent) choice of reward functions. To the best of our knowledge, our approach is the first to achieve optimal deployment complexity and optimal d dependence in sample complexity at the same time, even if the reward is known ahead of time. Our novel techniques include an exploration-preserving policy discretization and a generalized G-optimal experiment design, which could be of independent interest. Lastly, we analyze the related problem of regret minimization in low-adaptive RL and provide information-theoretic lower bounds for switching cost and batch complexity.

****

Provably efficient multi-task Reinforcement Learning in large state spaces

Baihe Huang,Jason D. Lee,Zhaoran Wang,Zhuoran Yang

We study multi-task Reinforcement Learning where shared knowledge among different environments is distilled to enable scalable generalization to a variety of problem instances. In the context of general function approximation, Markov Decision Process (MDP) with low Bilinear rank encapsulates a wide range of structural conditions that permit polynomial sample complexity in large state spaces, where the Bellman errors are related to bilinear forms of features with low intrinsic dimensions. To achieve multi-task learning in MDPs, we propose online representation learning algorithms to capture the shared features in the different task-specific bilinear forms. We show that in the presence of low-rank structures in the features of the bilinear forms, the algorithms benefit from sample complexity improvements compared to single-task learning. Therefore, we achieve the first sample efficient multi-task reinforcement learning algorithm with general function approximation.

****

An Equal-Size Hard EM Algorithm for Diverse Dialogue Generation

Yuqiao Wen,Yongchang Hao,Yanshuai Cao,Lili Mou

Open-domain dialogue systems aim to interact with humans through natural language texts in an open-ended fashion. Despite the recent success of super large dialogue systems such as ChatGPT, using medium-to-small-sized dialogue systems remains the common practice as they are more lightweight and accessible; however, generating diverse dialogue responses is challenging, especially with smaller models. In this work, we propose an Equal-size Hard Expectation--Maximization (EqHard-EM) algorithm to train a multi-decoder model for diverse dialogue generation. Our algorithm assigns a sample to a decoder in a hard manner and additionally imposes an equal-assignment constraint to ensure that all decoders are well-trained. We provide detailed theoretical analysis to justify our approach. Further, experiments on two large-scale open-domain dialogue datasets verify that our EqHard-EM algorithm generates high-quality diverse responses.

****

NeuralEQ: Neural-Network-Based Equalizer for High-Speed Wireline Communication

Hanseok Kim,Jae Hyung Ju,Hyun Seok Choi,Hyeri Roh,Woo-Seok Choi

Rapid growth of ML applications demand high-performance computing systems to perform massive data processing. In such systems, I/O bandwidth must be scaled up to prevent any performance degradation due to the limited data transfer rates. To meet this demand, recently wireline communication started adopting PAM4 signali ng and DSP-based equalizers. However, multi-level signaling and conventional equalizing techniques degrade the bit-error-rate (BER) performance significantly. To mitigate this problem, this paper proposes a novel neural network architecture that mimics the forward-backward algorithm estimating the posterior probabilities in Hidden Markov Models. The proposed neural network overcomes the existing equalizer performance such as feed-forward equalizers or decision-feedback equalizers, while reducing the complexity of the forward-backward algorithm.

****

Which is Better for Learning with Noisy Labels: The Semi-supervised Method or Modeling Label Noise?

Yu Yao,Mingming Gong,Yuxuan Du,Jun Yu,Bo Han,Kun Zhang,Tongliang Liu

In real life, accurately annotating large-scale datasets is sometimes difficult. Datasets used for training deep learning models are likely to contain label noise. To make use of the dataset containing label noise, two typical methods have been proposed. One is to employ the semi-supervised method by exploiting labeled \textit{confident examples} and unlabeled \textit{non-confident examples}. The other one is to \textit{model label noise} and design \textit{statistically consistent} classifiers. A natural question remains unsolved: which one should be used for a specific real-world application? In this paper, we answer the question from the perspective of \textit{causal data generative process}. Specifically, the semi-supervised method depends heavily on the data generation process while the modeling label noise method is independent of the generation process. For example, for a given dataset, if it has a causal generative structure that the features cause the label, the semi-supervised method would not be helpful. When the causal structure is unknown, we provide an intuitive method to discover the causal structure for a given dataset containing label noise.

****

The hidden uniform cluster prior in self-supervised learning

Mido Assran,Randall Balestriero,Quentin Duval,Florian Bordes,Ishan Misra,Piotr Bojanowski,Pascal Vincent,Michael Rabbat,Nicolas Ballas

A successful paradigm in representation learning is to perform self-supervised pretraining using tasks based on mini-batch statistics; (e.g., SimCLR, VICReg, SwAV, MSN). We show that in the formulation of all these methods is an overlooked prior to learn features that enable uniform clustering of the data. While this prior has led to remarkably semantic representations when pretraining on class-balanced data, such as ImageNet, we demonstrate that it can hamper performance when pretraining on class-imbalanced data. By moving away from conventional uniformity priors and instead preferring power-law distributed feature clusters, we show that one can improve the quality of the learned representations on real-world class-imbalanced datasets. To demonstrate this, we develop an extension of the Masked Siamese Networks (MSN) method to support the use of arbitrary features priors.

****

Revisiting Over-smoothing in Graph Neural Networks

Pantelis Elinas,Edwin V. Bonilla

Shallow graph neural networks (GNNs) are state-of-the-art models for relational data. However, it is known that deep GNNs suffer from over-smoothing where, as the number of layers increases, node representations become nearly indistinguishable and model performance on the downstream task degrades significantly. Despite multiple approaches being proposed to address this problem, it is unclear when any of these methods (or their combination) works best and how they perform when evaluated under exactly the same experimental setting. In this paper, we systematically and carefully evaluate different methods for alleviating over-smoothing in GNNs. Furthermore, inspired by standard deeply supervised nets, we propose a general architecture that helps alleviate over-smoothing based on the idea of layer-wise supervision. We term this architecture deeply supervised GNNs (or DSGNNs for short). Our experiments show that deeper GNNs can indeed provide better performance when trained on a combination of different approaches and that DSGNNs are robust under various conditions and can provide the best performance in missing-feature scenarios.

****

Mosaic Representation Learning for Self-supervised Visual Pre-training

Zhaoqing Wang,Ziyu Chen,Yaqian Li,Yandong Guo,Jun Yu,Mingming Gong,Tongliang Liu

Self-supervised learning has achieved significant success in learning visual representations without the need for manual annotation. To obtain generalizable representations, a meticulously designed data augmentation strategy is one of the most crucial parts. Recently, multi-crop strategies utilizing a set of small crops as positive samples have been shown to learn spatially structured features. However, it overlooks the diverse contextual backgrounds, which reduces the variance of the input views and degenerates the performance. To address this problem, we propose a mosaic representation learning framework (MosRep), consisting of a new data augmentation strategy that enriches the backgrounds of each small crop and improves the quality of visual representations. Specifically, we randomly sample numbers of small crops from different input images and compose them into a mosaic view, which is equivalent to introducing different background information for each small crop. Additionally, we further jitter the mosaic view to prevent memorizing the spatial locations of each crop. Along with optimization, our MosRep gradually extracts more discriminative features. Extensive experimental results demonstrate that our method improves the performance far greater than the multi-crop strategy on a series of downstream tasks, e.g., +7.4% and +4.9% than the multi-crop strategy on ImageNet-1K with 1% label and 10% label, respectively. Code is available at https://github.com/DerrickWang005/MosRep.git.

****

FluidLab: A Differentiable Environment for Benchmarking Complex Fluid Manipulation

Zhou Xian,Bo Zhu,Zhenjia Xu,Hsiao-Yu Tung,Antonio Torralba,Katerina Fragkiadaki,Chuang Gan

Humans manipulate various kinds of fluids in their everyday life: creating latte art, scooping floating objects from water, rolling an ice cream cone, etc. Using robots to augment or replace human labors in these daily settings remain as a challenging task due to the multifaceted complexities of fluids. Previous research in robotic fluid manipulation mostly consider fluids governed by an ideal, Newtonian model in simple task settings (e.g., pouring water into a container). However, the vast majority of real-world fluid systems manifest their complexities in terms of the fluid’s complex material behaviors (e.g., elastoplastic deformation) and multi-component interactions (e.g. coffee and frothed milk when making latte art), both of which were well beyond the scope of the current literature. To evaluate robot learning algorithms on understanding and interacting with such complex fluid systems, a comprehensive virtual platform with versatile simulation capabilities and well-established tasks is needed. In this work, we introduce FluidLab, a simulation environment with a diverse set of manipulation tasks involving complex fluid dynamics. These tasks address interactions between solid and fluid as well as among multiple fluids. At the heart of our platform is a fully differentiable physics simulator, FluidEngine, providing GPU-accelerated simulations and gradient calculations for various material types and their couplings, extending the scope of the existing differentiable simulation engines. We identify several challenges for fluid manipulation learning by evaluating a set of reinforcement learning and trajectory optimization methods on our platform. To address these challenges, we propose several domain-specific optimization schemes coupled with differentiable physics, which are empirically shown to be effective in tackling optimization problems featured by fluid system’s non-convex and non-smooth properties. Furthermore, we demonstrate reasonable sim-to-real transfer by deploying optimized trajectories in real-world settings. FluidLab is publicly available at: https://fluidlab2023.github.io.

****

Route, Interpret, Repeat: Blurring the Line Between Posthoc Explainability and Interpretable Models

Shantanu Ghosh,Ke Yu,Forough Arabshahi,kayhan Batmanghelich

The current approach to ML model design is either to choose a flexible Blackbox model and explain it post hoc or to start with an interpretable model. Blackbox models are flexible but difficult to explain, whereas interpretable models are designed to be explainable. However, developing interpretable models necessitates extensive ML knowledge, and the resulting models tend to be less flexible, offering potentially subpar performance compared to their Blackbox equivalents. This paper aims to blur the distinction between a post hoc explanation of a BlackBox and constructing interpretable models. We propose beginning with a flexible BlackBox model and gradually carving out a mixture of interpretable models and a residual network. Our design identifies a subset of samples and routes them through the interpretable models. The remaining samples are routed through a flexible residual network. We adopt First Order Logic (FOL) as the interpretable model's backbone, which provides basic reasoning on the concept retrieved from the BlackBox model. On the residual network, we repeat the method until the proportion of data explained by the residual network falls below a desired threshold.

Our approach offers several advantages. First, the mixture of interpretable and flexible residual networks results in almost no compromise in performance. Second, the rout, interpret, and repeat approach yields a highly flexible interpretable model. Our extensive experiment demonstrates the performance of the model on various datasets. We show that by editing the FOL model, we can fix the shortcut learned by the original BlackBox model. Finally, our method provides a framework for a hybrid symbolic-connectionist network that is simple to train and adaptable to many applications.

****

On Regularization for Explaining Graph Neural Networks: An Information Theory Perspective

Junfeng Fang,Wei Liu,An Zhang,Xiang Wang,Xiangnan He,Kun Wang,Tat-Seng Chua

This work studies the explainability of graph neural networks (GNNs), which is important for the credibility of GNNs in practical usage. Existing work mostly follows the two-phase paradigm to interpret a prediction: feature attribution and selection. However, another important component --- regularization, which is crucial to facilitate the above paradigm --- has been seldom studied. In this work, we explore the role of regularization in GNNs explainability from the perspective of information theory. Our main findings are: 1) regularization is essentially pursuing the balance between two phases, 2) its optimal coefficient is proportional to the sparsity of explanations, 3) existing methods imply an implicit regularization effect of stochastic mechanism, and 4) its contradictory effects on two phases are responsible for the out-of-distribution (OOD) issue in post-hoc explainability. Based on these findings, we propose two common optimization methods, which can bolster the performance of the current explanation methods via sparsity-adaptive and OOD-resistant regularization schemes. Extensive empirical studies validate our findings and proposed methods. Code is available at https://anonymous.4open.science/r/Rethink_Reg-07F0.

****

The Dark Side of Invariance: Revisiting the Role of Augmentations in Contrastive Learning

Alex Tamkin,Margalit Glasgow,Xiluo He,Noah Goodman

What role do augmentations play in contrastive learning? Recent work suggests that good augmentations are label-preserving with respect to a specific downstream task. We complicate this picture by showing that label-destroying augmentations are often crucial in the foundation model setting, where the goal is to learn diverse, general-purpose representations for multiple downstream tasks. We perform contrastive learning experiments on a range of image and audio datasets with multiple downstream tasks (e.g. for digits superimposed on photographs, predicting the class of one vs. the other). We find that Viewmaker Networks, a recently proposed model for learning augmentations for contrastive learning, produce label-destroying augmentations that stochastically destroy features needed for different downstream tasks. These augmentations are interpretable (e.g. altering shapes, digits, or letters added to images) and surprisingly often result in better performance compared to expert-designed augmentations, despite not preserving lab el information. To support our empirical results, we theoretically analyze a simple contrastive learning setting with a linear model. In this setting, label-destroying augmentations are crucial for preventing one set of features from suppressing the learning of features useful for another downstream task. Our results highlight the need for analyzing the interaction between multiple downstream tasks when trying to explain the success of foundation models

****

Hierarchical Gaussian Mixture based Task Generative Model for Robust Meta-Learning

Yizhou Zhang,Jingchao Ni,Wei Cheng,Zhengzhang Chen,Liang Tong,Haifeng Chen

Meta-learning enables quick adaptation of machine learning models to new tasks with limited data. While tasks could come from varying distributions in reality, most of the existing meta-learning methods consider both training and testing tasks as from the same uni-component distribution, overlooking two critical needs of a practical solution: (1) the various sources of tasks may compose a multi-component mixture distribution, and (2) novel tasks may come from a distribution that is unseen during meta-training. In this paper, we demonstrate these two challenges can be solved jointly by modeling the density of task instances. We develop a meta-training framework underlain by a novel Hierarchical Gaussian Mixture based Task Generative Model (HTGM). HTGM extends the widely used empirical process of sampling tasks to a theoretical model, which learns task embeddings, fits mixture distribution of tasks, and enables density-based scoring of novel tasks. The framework is agnostic to the encoder and scales well with large backbone networks. The model parameters are learned end-to-end by maximum likelihood estimation via an Expectation-Maximization algorithm. Extensive experiments on benchmark datasets indicate the effectiveness of our method for both sample classification and novel task detection.

****

Game-Theoretic Understanding of Misclassification

Kosuke Sumiyasu,Kazuhiko Kawamoto,Kera Hiroshi

This paper analyzes various types of image misclassification from a game-theoretic view. Particularly, we consider the misclassification of clean, adversarial, and corrupted images and characterize it through the distribution of multi-order interactions. We discover that the distribution of multi-order interactions varies across the types of misclassification. For example, misclassified adversarial images have a higher strength of high-order interactions than correctly classified clean images, which indicates that adversarial perturbations create spurious features that arise from complex cooperation between pixels. By contrast, misclassified corrupted images have a lower strength of low-order interactions than correctly classified clean images, which indicates that corruptions break the local cooperation between pixels. We also provide the first analysis of Vision Transformers using interactions. We found that Vision Transformers show a different tendency in the distribution of interactions from that in CNNs, and this implies that they exploit the features that CNNs do not use for the prediction. Our study demonstrates that the recent game-theoretic analysis of deep learning models can be broadened to analyze various malfunctions of deep learning models including Vision Transformers by using the distribution, order, and sign of interactions.

****

The Final Ascent: When Bigger Models Generalize Worse on Noisy-Labeled Data

Yihao Xue,Kyle Whitecross,Baharan Mirzasoleiman

Increasing the size of overparameterized neural networks has been shown to improve their generalization performance. However, real-world datasets often contain a significant fraction of noisy labels, which can drastically harm the performance of the models trained on them. In this work, we study how neural networks' test loss changes with model size when the training set contains noisy labels. We show that under a sufficiently large noise-to-sample size ratio, generalization error eventually increases with model size. First, we provide a theoretical analysis on random feature regression and show that this phenomenon occurs as the variance of the generalization loss experiences a second ascent under large noise-to-sample size ratio. Then, we present extensive empirical evidence confirming that our theoretical results hold for neural networks. Furthermore, we empirically observe that the adverse effect of network size is more pronounced when robust training methods are employed to learn from noisy-labeled data. Our results have important practical implications: First, larger models should be employed with extra care, particularly when trained on smaller dataset or using robust learning methods. Second, a large sample size can alleviate the effect of noisy labels and allow larger models to achieve a superior performance even under noise.

****

Long-Tailed Partial Label Learning via Dynamic Rebalancing

Feng Hong,Jiangchao Yao,Zhihan Zhou,Ya Zhang,Yanfeng Wang

Real-world data usually couples the label ambiguity and heavy imbalance, challenging the algorithmic robustness of partial label learning (PLL) and long-tailed learning (LT). The straightforward combination of LT and PLL, i.e., LT-PLL, suffers from a fundamental dilemma: LT methods build upon a given class distribution that is unavailable in PLL, and the performance of PLL is severely influenced in long-tailed context. We show that even with the auxiliary of an oracle class prior, the state-of-the-art methods underperform due to an adverse fact that the constant rebalancing in LT is harsh to the label disambiguation in PLL. To overcome this challenge, we thus propose a dynamic rebalancing method, termed as RECORDS, without assuming any prior knowledge about the class distribution. Based on a parametric decomposition of the biased output, our method constructs a dynamic adjustment that is benign to the label disambiguation process and theoretically converges to the oracle class prior. Extensive experiments on three benchmark datasets demonstrate the significant gain of RECORDS compared with a range of baselines. The code is publicly available.

****

Task Ambiguity in Humans and Language Models

Alex Tamkin,Kunal Handa,Avash Shrestha,Noah Goodman

Language models have recently achieved strong performance across a wide range of NLP benchmarks. However, real world tasks are often poorly specified, and agents must deduce the intended behavior from a combination of context, instructions, and examples. We investigate how both humans and models behave in the face of such task ambiguity by proposing AmbiBench, a new benchmark of six ambiguously-specified classification tasks. We evaluate humans and models on AmbiBench by seeing how well they identify the intended task using 1) instructions with varying degrees of ambiguity, and 2) different numbers of labeled examples. We find that the combination of model scaling (to 175B parameters) and reinforcement learning from human feedback (RLHF) enables models to approach or exceed the accuracy of human participants across tasks, but that either one of these alone is not sufficient. In addition, we show how to dramatically improve the accuracy of language models trained without RLHF by finetuning on a small number of ambiguous in-context examples, providing a promising direction for teaching models to generalize well in the face of ambiguity.

****

Equivariant Disentangled Transformation for Domain Generalization under Combination Shift

Yivan Zhang,Jindong Wang,Xing Xie,Masashi Sugiyama

Machine learning systems may encounter unexpected problems when the data distribution changes in the deployment environment. A major reason is that certain combinations of domains and labels are not observed during training but appear in the test environment. Although various invariance-based algorithms can be applied, we find that the performance gain is often marginal. To formally analyze this issue, we provide a unique algebraic formulation of the combination shift problem based on the concepts of homomorphism, equivariance, and a refined definition of disentanglement. The algebraic requirements naturally derive a simple yet effective method, referred to as equivariant disentangled transformation (EDT), which augments the data based on the algebraic structures of labels and makes the transformation satisfy the equivariance and disentanglement requirements. Experimental results demonstrate that invariance may be insufficient, and it is importan

t to exploit the equivariance structure in the combination shift problem.
****

Best Possible Q-Learning

Jiechuan Jiang,Zongqing Lu

Fully decentralized learning, where the global information, i.e., the actions of other agents, is inaccessible, is a fundamental challenge in multi-agent reinforcement learning. However, the convergence and optimality of most decentralized algorithms are not theoretically guaranteed, since the transition probabilities are non-stationary as all agents are updating policies simultaneously. To tackle this challenge, we propose \textit{best possible operator}, a novel decentralized operator, and prove that the policies of agents will converge to the optimal joint policy if each agent independently updates its individual state-action value by the operator. Further, to make the update more efficient and practical, we simplify the operator and prove that the convergence and optimality still hold with the simplified one. By instantiating the simplified operator, the derived fully decentralized algorithm, best possible Q-learning (BQL), does not suffer from non-stationarity. Empirically, we show that BQL achieves remarkable improvement over baselines in a variety of cooperative multi-agent tasks.
****

Winning Both the Accuracy of Floating Point Activation and the Simplicity of Integer Arithmetic

Yulhwa Kim,Jaeyong Jang,Jehun Lee,Jihoon Park,Jeonghoon Kim,Byeongwook Kim,Baeseong park,Se Jung Kwon,Dongsoo Lee,jae-joon kim

Even though floating point (FP) numbers have been adopted as a de facto standard data format for deep learning computing, the complexity of FP arithmetic impedes a broader deployment of Deep Neural Networks (DNNs). Recent works such as quantization have attempted to replace the FP matrix multiplication (MatMul) of DNNs with simple integer MatMul by transforming the datatypes of both weights and activations into integers. Unfortunately, unlike weight values that are static, it is challenging to represent dynamic activations with integers. In this paper, to simultaneously achieve the accuracy of FP activation and the simplicity of integer arithmetic, we present a method for replacing FP arithmetic with integer one without changing FP activations in the storage format while weights are quantized. The proposed method pre-aligns the significands of FP activations just ahead of the MatMul on-the-fly so that the aligned significands (integers) can be used for the computation. Inspired by an observation that conventional FP arithmetic does not produce precise results due to rounding, we demonstrate that our proposed integer arithmetic-based scheme can produce the same level of errors as that of the FP arithmetic in case DNNs use FP activations and quantized weights. Experimental results show that the hardware based on the proposed scheme shows significant improvement over FP arithmetic-based designs in terms of energy efficiency and throughput-per-area while maintaining a similar level of accuracy.
****

Preference Transformer: Modeling Human Preferences using Transformers for RL

Changyeon Kim,Jongjin Park,Jinwoo Shin,Honglak Lee,Pieter Abbeel,Kimin Lee

Preference-based reinforcement learning (RL) provides a framework to train agents using human preferences between two behaviors. However, preference-based RL has been challenging to scale since it requires a large amount of human feedback to learn a reward function aligned with human intent. In this paper, we present Preference Transformer, a neural architecture that models human preferences using transformers. Unlike prior approaches assuming human judgment is based on the Markovian rewards which contribute to the decision equally, we introduce a new preference model based on the weighted sum of non-Markovian rewards. We then design the proposed preference model using a transformer architecture that stacks causal and bidirectional self-attention layers. We demonstrate that Preference Transformer can solve a variety of control tasks using real human preferences, while prior approaches fail to work. We also show that Preference Transformer can induce a well-specified reward and attend to critical events in the trajectory by automatically capturing the temporal dependencies in human decision-making. Code is available on the project website: https://sites.google.com/view/preference-tr ansformer.

****

Flow Matching for Generative Modeling

Yaron Lipman,Ricky T. Q. Chen,Heli Ben-Hamu,Maximilian Nickel,Matthew Le

We introduce a new paradigm for generative modeling built on Continuous Normalizing Flows (CNFs), allowing us to train CNFs at unprecedented scale. Specifically, we present the notion of Flow Matching (FM), a simulation-free approach for training CNFs based on regressing vector fields of fixed conditional probability paths. Flow Matching is compatible with a general family of Gaussian probability paths for transforming between noise and data samples---which subsumes existing diffusion paths as specific instances. Interestingly, we find that employing FM with diffusion paths results in a more robust and stable alternative for training diffusion models. Furthermore, Flow Matching opens the door to training CNFs with other, non-diffusion probability paths. An instance of particular interest is using Optimal Transport (OT) displacement interpolation to define the conditional probability paths. These paths are more efficient than diffusion paths, provide faster training and sampling, and result in better generalization. Training CNFs using Flow Matching on ImageNet leads to consistently better performance than alternative diffusion-based methods in terms of both likelihood and sample quality, and allows fast and reliable sample generation using off-the-shelf numerical ODE solvers.

****

Graph-informed Neural Point Process With Monotonic Nets

Zhimeng Pan,Zheng Wang,Shandian Zhe

Multi-class event data is ubiquitous in real-world applications. The recent neural temporal point processes have used monotonic nets to model the cumulative conditional intensity to avoid an intractable integration in the likelihood. While successful, they are restricted to single-type events and easily sink into poor learning results. To address these limitations and exploit valuable structural information within event participants, we develop a Graph-Informed Neural Point Process (GINPP) that can freely handle multiple event types, greatly improve learning efficiency with the monotonic net, and effectively integrate the graph information to facilitate training. First, we find the bottleneck of the previous model arises from the standard soft-plus transformation over the output of the monotonic net, which greatly enlarges the prediction variations of the monotonic net and increases the training challenge. We propose a shift-scale version that can significantly reduce the variation and promote learning efficiency. Second, we use a conditional mark distribution to model multiple event types, without the need for explicitly estimating the intensity for each type. The latter can be much more challenging. Third, we use random walks to collect the neighborhood of each event participant and use an attention mechanism to update the hidden state of each participant according to the observed events of both the participant itself and its neighborhood. In this way, we can effectively leverage the graph knowledge, and scale up to large graphs. We have shown the advantage of our approach in both ablation studies and real-world applications.

****

Targeted Hyperparameter Optimization with Lexicographic Preferences Over Multiple Objectives

Shaokun Zhang,Feiran Jia,Chi Wang,Qingyun Wu

Motivated by various practical applications, we propose a novel and general formulation of targeted multi-objective hyperparameter optimization. Our formulation allows a clear specification of an automatable optimization goal using lexicographic preference over multiple objectives. We then propose a randomized directed search method named LexiFlow to solve this problem. We demonstrate the strong empirical performance of the proposed algorithm in multiple hyperparameter optimization tasks.

****

Learning to Decouple Complex System for Sequential Data

Zihan Zhou,Minxin Ao,Tianshu Yu

A complex system with cluttered observations may be a coupled mixture of multipl e simple sub-systems corresponding to \emph{latent entities}. Such sub-systems may hold distinct dynamics in the continuous-time domain, therein complicated interactions between sub-systems also evolve over time. This setting is fairly common in the real world, but has been less considered. In this paper, we propose a sequential learning approach under this setting by decoupling a complex system for handling irregularly sampled and cluttered sequential observations. Such decoupling brings about not only subsystems describing the dynamics of each latent entity, but also a meta-system capturing the interaction between entities over time. Specifically, we argue that the meta-system of interactions is governed by a smoothed version of \emph{projected differential equations}. Experimental results on synthetic and real-world datasets show the advantages of our approach when facing complex and cluttered sequential data compared to the state-of-the-art.

****

Restoration based Generative Models

Jaemoo Choi,Yesom Park,Myungjoo Kang

Denoising generative models (DGMs) have recently attracted increasing attention by showing impressive synthesis quality. DGMs are built on a diffusion process that pushes data to the noise distribution and the models learn to denoise. In this paper, we establish the interpretation of DGMs in terms of image restoration (IR). Integrating IR literature allows us to use an alternative objective and diverse forward processes, not confining to the diffusion process. By imposing prior knowledge on the loss function grounded on MAP estimation, we eliminate the need for the expensive sampling of DGMs. Also, we propose a multi-scale training, which alleviates the latent inefficiency of DGMs, by taking advantage of the flexibility of the forward process. Our model improves the quality and efficiency of both training and inference, achieving state-of-the-art performance when the number of forward steps is limited. Furthermore, we show the applicability of our model to inverse problems. We believe that our framework paves the way for designing a new type of flexible general generative model.

****

How hard are computer vision datasets? Calibrating dataset difficulty to viewing time

David Mayo,Jesse Cummings,Xinyu Lin,Dan Gutfreund,Boris Katz,Andrei Barbu

Humans outperform object recognizers despite the fact that models perform well on current datasets. Numerous attempts have been made to create more challenging datasets by scaling them up from the web, exploring distribution shift, or adding controls for biases. The difficulty of each image in each dataset is not independently evaluated, nor is the concept of dataset difficulty as a whole well-posed. We develop a new dataset difficulty metric based on how long humans must view an image in order to classify a target object. Images whose objects can be recognized in 17ms are considered to be easier than those which require seconds of viewing time. Using 133,588 judgments on two major datasets, ImageNet and ObjectNet, we determine the distribution of image difficulties in those datasets, which we find varies wildly, but significantly undersamples hard images. Rather than hoping that distribution shift or other approaches will lead to hard datasets, we should measure the difficulty of datasets and seek to explicitly fill out the class of difficult examples. Analyzing model performance guided by image difficulty reveals that models tend to have lower performance and a larger generalization gap on harder images. Encouragingly for the biological validity of current architectures, much of the variance in human difficulty can be accounted for given an object recognizer by computing a combination of prediction depth, c-score, and adversarial robustness. We release a dataset of such judgments as a complementary metric to raw performance and a network’s ability to explain neural recordings. Such experiments with humans allow us to create a metric for progress in object recognition datasets, which we find are skewed toward easy examples, to test the biological validity of models in a novel way, and to develop tools for shaping datasets as they are being gathered to focus them on filling out the missing class of hard examples from today’s datasets. Dataset and analysis code can be found at https://github.com/image-flash/image-flash-2022.

****

Self-Supervised Logit Adjustment

Zhihan Zhou,Jiangchao Yao,Feng Hong,Yanfeng Wang,Bo Han,Ya Zhang

Self-supervised learning (SSL) has achieved tremendous success on various well curated datasets in computer vision and natural language processing. Nevertheless, it is hard for existing works to capture transferable and robust features, when facing the long-tailed distribution in the real-world scenarios. The attribution is that plain SSL methods to pursue sample-level uniformity easily leads to the distorted embedding space, where head classes with the huge sample number dominate the feature regime and tail classes passively collapse. To tackle this problem, we propose a novel Self-Supervised Logit Adjustment (S^2LA) method to achieve the category-level uniformity from a geometric perspective. Specially, we measure the geometric statistics of the embedding space to construct the calibration, and jointly learn a surrogate label allocation to constrain the space expansion of head classes and avoid the passive collapse of tail classes. Our proposal does not alter the setting of SSL and can be easily integrated into existing works in an end-to-end and low-cost manner. Extensive results on a range of benchmark datasets show the effectiveness of S^2LA with high tolerance to the distribution skewness.

****

Proportional Amplitude Spectrum Training Augmentation for Synthetic-to-Real Domain Generalization

Prithvijit Chattopadhyay,Kartik Sarangmath,Vivek Vijaykumar,Judy Hoffman

Synthetic data offers the promise of cheap and bountiful training data for settings where lots of labeled real-world data for some task is unavailable. However, models trained on synthetic data significantly underperform on real-world data. In this paper, we propose Proportional Amplitude Spectrum Training Augmentation (PASTA), a simple and effective augmentation strategy to improve out-of-the-box synthetic-to-real (syn-to-real) generalization performance. PASTA involves perturbing the amplitude spectrums of the synthetic images in the Fourier domain to generate augmented views. We design PASTA to perturb the amplitude spectrums in a structured manner such that high-frequency components are perturbed relatively more than the low-frequency ones. For the tasks of semantic segmentation (GTAV→Real), object detection (Sim10K→Real), and object recognition (VisDAC Syn→Real), across a total of 5 syn-to-real shifts, we find that PASTA either outperforms or is consistently competitive with more complex state-of-the-art methods while being complementary to other generalization approaches.

****

More Centralized Training, Still Decentralized Execution: Multi-Agent Conditional Policy Factorization

Jiangxing Wang,Deheng Ye,Zongqing Lu

In cooperative multi-agent reinforcement learning (MARL), combining value decomposition with actor-critic enables agents to learn stochastic policies, which are more suitable for the partially observable environment. Given the goal of learning local policies that enable decentralized execution, agents are commonly assumed to be independent of each other, even in centralized training. However, such an assumption may prohibit agents from learning the optimal joint policy. To address this problem, we explicitly take the dependency among agents into centralized training. Although this leads to the optimal joint policy, it may not be factorized for decentralized execution. Nevertheless, we theoretically show that from such a joint policy, we can always derive another joint policy that achieves the same optimality but can be factorized for decentralized execution. To this end, we propose multi-agent conditional policy factorization (MACPF), which takes more centralized training but still enables decentralized execution. We empirically verify MACPF in various cooperative MARL tasks and demonstrate that MACPF achieves better performance or faster convergence than baselines. Our code is available at https://github.com/PKU-RL/FOP-DMAC-MACPF.

****

GAPS: Few-Shot Incremental Semantic Segmentation via Guided Copy-Paste Synthesis

Ri-Zhao Qiu,Peiyi Chen,Wangzhe Sun,Yu-Xiong Wang,Kris Hauser

Few-shot incremental segmentation is the task of updating a segmentation model, as novel classes are introduced online overtime with a small number of training images. Although incremental segmentation methods exist in the literature, they tend to fall short in the few-shot regime and when given partially-annotated training images, where only the novel class is segmented. This paper proposes a data synthesizer, Guided copy-And-Paste Synthesis (GAPS), that improves the performance of few-shot incremental segmentation in a model-agnostic fashion. Despite the great success of copy-paste synthesis in the conventional offline visual recognition, we demonstrate substantially degraded performance of its naive extension in our online scenario, due to newly encountered challenges. To this end, GAPS (i) addresses the partial-annotation problem by leveraging copy-paste to generate fully-labeled data for training, (ii) helps augment the few images of novel objects by introducing a guided sampling process, and (iii) mitigates catastrophic forgetting by employing a diverse memory-replay buffer. Compared to existing state-of-the-art methods, GAPS dramatically boosts the novel IoU of baseline methods on established few-shot incremental segmentation benchmarks by up to 80%. More notably, GAPS maintains good performance in even more impoverished annotation settings, where only single instances of novel objects are annotated.

****

Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks

Bowen Jin,Yu Zhang,Yu Meng,Jiawei Han

Edges in many real-world social/information networks are associated with rich text information (e.g., user-user communications or user-product reviews). However, mainstream network representation learning models focus on propagating and aggregating node attributes, lacking specific designs to utilize text semantics on edges. While there exist edge-aware graph neural networks, they directly initialize edge attributes as a feature vector, which cannot fully capture the contextualized text semantics of edges. In this paper, we propose Edgeformers, a framework built upon graph-enhanced Transformers, to perform edge and node representation learning by modeling texts on edges in a contextualized way. Specifically, in edge representation learning, we inject network information into each Transformer layer when encoding edge texts; in node representation learning, we aggregate edge representations through an attention mechanism within each node’s ego-graph. On five public datasets from three different domains, Edgeformers consistently outperform state-of-the-art baselines in edge classification and link prediction, demonstrating the efficacy in learning edge and node representations, respectively.

****

How Distinguishable Are Vocoder Models? Analyzing Vocoder Fingerprints for Fake Audio

Chu Yuan Zhang,Jiangyan Yi,Jianhua Tao,Tao Wang,Xinrui Yan,Chenglong Wang

In recent years, vocoders powered by deep neural networks (DNNs) have found much success in the task of generating raw waveforms from acoustic features, as the audio generated becomes increasingly realistic. This development however raises a few challenges, especially in the field of forensics, where the attribution of audio to real or generated sources is vital. To our knowledge, our investigation constitutes the first efforts to answer the question on the existence of vocoder fingerprints and to analyze them. In this paper, we present our discoveries in identifying the sources of generated audio waveforms. Our experiments conducted on the multi-speaker LibriTTS dataset show that (1) vocoder models do leave model-specific fingerprints on the audio they generate, and (2) minor differences in vocoder training can result in sufficiently different fingerprints in generated audio as to allow for distinguishing between the two. We believe that these differences are strong evidence that there exist vocoder-specific fingerprints that can be exploited for source identification purposes.

****

Any-scale Balanced Samplers for Discrete Space

Haoran Sun,Bo Dai,Charles Sutton,Dale Schuurmans,Hanjun Dai

The locally balanced informed proposal has proved to be highly effective for sampling from discrete spaces. However, its success relies on the "local'' factor, which ensures that whenever the proposal distribution is restricted to be near the current state, the locally balanced weight functions are asymptotically optimal and the gradient approximations are accurate. In seeking a more efficient sampling algorithm, many recent works have considered increasing the scale of the proposal distributions, but this causes the "local'' factor to no longer hold. Instead, we propose any-scale balanced samplers to repair the gap in non-local proposals. In particular, we substitute the locally balanced function with an any-scale balanced function that can self-adjust to achieve better efficiency for proposal distributions at any scale. We also use quadratic approximations to capture curvature of the target distribution and reduce the error in the gradient approximation, while employing a Gaussian integral trick with a special estimated diagonal to efficiently sample from the quadratic proposal distribution. On various synthetic and real distributions, the proposed sampler substantially outperforms existing approaches.

****

BinSGDM: Extreme One-Bit Quantization for Communication Efficient Large-Scale Distributed Training

Hanyang Peng,Shuang Qin,Yue Yu,Jin Wang,Hui Wang,Ge Li

To alleviate the communication bottleneck of large-scale distributed training, a rich body of prior communication-compression optimizers have been proposed. These methods focus mainly on high compression ratio to expect acceleration. However, some recent works pointed out, when running with distributed training frameworks (\emph{e.g.}, \emph{DistributedDataParallel} in pytorch), these methods may provide no acceleration over the off-the-shelve uncompressed SGD/Adam in the typical settings, due to heavy compression/decompression computation or incompatibility with efficient communication primitives or the requirement of uncompressed warmup at the early stage. For these reasons, we propose a novel extreme one-bit quantization optimizer, dubbed \emph{BinSGDM}. The quantization of \emph{BinSGDM} is computed easily and lightly, and it does not need to resort to uncompressed optimizers for warmup. We also theoretically prove that it can promise the same convergence speed as the original Adam. Moreover, we specially present a hierarchical communication scheme to further lower the communication volume. Extensive experiments are conducted on 8 to 64 GPUs (1 to 8 nodes) for distributed training with \emph{DistributedDataParallel}, and the experimental results demonstrates that \emph{BinSGDM} with the communication scheme can achieve up to {$\bm{2.47 \times}$} speedup for training ResNet-50 and $\bm{6.26\times}$ speedup for training BERT-Base, compared to the full-precision optimizers.

****

Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design

Keir Adams,Connor W. Coley

Shape-based virtual screening is widely used in ligand-based drug design to search chemical libraries for molecules with similar 3D shapes yet novel 2D graph structures compared to known ligands. 3D deep generative models can potentially automate this exploration of shape-conditioned 3D chemical space; however, no existing models can reliably generate geometrically realistic drug-like molecules in conformations with a specific shape. We introduce a new multimodal 3D generative model that enables shape-conditioned 3D molecular design by equivariantly encoding molecular shape and variationally encoding chemical identity. We ensure local geometric and chemical validity of generated molecules by using autoregressive fragment-based generation with heuristic bonding geometries, allowing the model to prioritize the scoring of rotatable bonds to best align the growing conformation to the target shape. We evaluate our 3D generative model in tasks relevant to drug design including shape-conditioned generation of chemically diverse molecular structures and shape-constrained molecular property optimization, demonstrating its utility over virtual screening of enumerated libraries.

****

Leaves: Learning Views for Time-Series Data in Contrastive Learning

Han Yu,Huiyuan Yang,Akane Sano

Contrastive learning, a self-supervised learning method that can learn represent ations from unlabeled data, has been developed promisingly. Many methods of contrastive learning depend on data augmentation techniques, which generate different views from the original signal. However, tuning policies and hyper-parameters for more effective data augmentation methods in contrastive learning is often time and resource-consuming. Researchers have designed approaches to automatically generate new views for some input signals, especially on the image data. But the view-learning method is not well developed for time-series data. In this work, we propose a simple but effective module for automating view generation for time-series data in contrastive learning, named learning views for time-series data (LEAVES). The proposed module learns the hyper-parameters for augmentations using adversarial training in contrastive learning. We validate the effectiveness of the proposed method using multiple time-series datasets. The experiments demonstrate that the proposed method is more effective in finding reasonable views and performs downstream tasks better than the baselines, including manually tuned augmentation-based contrastive learning methods and SOTA methods.

****

The Eigenlearning Framework: A Conservation Law Perspective on Kernel Ridge Regression and Wide Neural Networks

James B Simon,Madeline Dickens,Dhruva Karkada,Michael Deweese

We derive simple closed-form estimates for the test risk and other generalization metrics of kernel ridge regression (KRR). Relative to prior work, our derivations are greatly simplified and our final expressions are far more interpretable. These improvements are enabled by our identification of a sharp conservation law which limits the ability of KRR to learn any orthonormal basis of functions. Test risk and other objects of interest are expressed in a transparent, interpretable way in terms of our conserved quantity evaluated in the kernel eigenbasis. We use our improved framework to:

i) provide a theoretical explanation for the ``deep bootstrap" of Nakkiran et al (2020),

ii) prove a new result regarding the hardness of the classic parity problem,

iii) fashion a theoretical tool for the study of adversarial robustness, and

iv) draw a tight analogy between KRR and a well-studied system in statistical physics.

****

Imbalanced Semi-supervised Learning with Bias Adaptive Classifier

Renzhen Wang,Xixi Jia,Quanziang Wang,Yichen Wu,Deyu Meng

Pseudo-labeling has proven to be a promising semi-supervised learning (SSL) paradigm. Existing pseudo-labeling methods commonly assume that the class distributions of training data are balanced. However, such an assumption is far from realistic scenarios and thus severely limits the performance of current pseudo-labeling methods under the context of class-imbalance. To alleviate this problem, we design a bias adaptive classifier that targets the imbalanced SSL setups. The core idea is to automatically assimilate the training bias caused by class imbalance via the bias adaptive classifier, which is composed of a novel bias attractor and the original linear classifier. The bias attractor is designed as a light-weight residual network and learned through a bi-level learning framework, which enables the bias adaptive classifier to fit imbalanced training data, while the linear classifier can provide unbiased label prediction for each class. We conduct extensive experiments under various imbalanced semi-supervised setups, and the results demonstrate that our method can be applied to different pseudo-labeling models and is superior to current state-of-the-art methods.

****

COMNET : CORTICAL MODULES ARE POWERFUL

Ashish Kumar,Laxmidhar Behera

Existing CNN architectures may achieve efficiency in either one or two dimensions: FLOPs, depth, accuracy, representation power, latency but not in all. In this work, we present a pragmatically designed novel CNN architecture “CoMNet” which offers multi-dimensional efficiency at once such as: simple yet accurate, lower latency and FLOPs, high representation power in limited parameters, low memory consumption, negligible branching, smaller depths, and only a few design hyperpa rameters. The key to achieve the multi-dimensional efficiency is our use of biological underpinnings into CoMNet which is primarily the organization of cortical modules in the visual cortex. To realize CoMNet, a few concepts from well understood CNN designs are directly inherited such as residual learning. Our solid experimental evaluations demonstrate superiority of CoMNet over many state-of-the-art industry and academia dominant architectures such as ResNet, RepVGG etc. For instance, CoMNet supersedes ResNet-50 on ImageNet while being 50% shallower, 22% lesser parameters, 25% lower FLOPs and latency, and in 16% lesser training epochs. Code will be opensourced post the reviews.

****

Do We Always Need to Penalize Variance of Losses for Learning with Label Noise?

Yexiong Lin,Yu Yao,Yuxuan Du,Jun Yu,Bo Han,Mingming Gong,Tongliang Liu

Algorithms which minimize the averaged loss have been widely designed for dealing with noisy labels. Intuitively, when there is a finite training sample, penalizing the variance of losses will improve the stability and generalization of the algorithms. Interestingly, we found that the variance of losses sometimes needs to be increased for the problem of learning with noisy labels. Specifically, increasing the variance of losses would boost the memorization effect and reduce the harmfulness of incorrect labels. Regularizers can be easily designed to increase the variance of losses and be plugged in many existing algorithms. Empirically, the proposed method by increasing the variance of losses could improve the generalization ability of baselines on both synthetic and real-world datasets.

****

DeepGuiser: Learning to Disguise Neural Architectures for Impeding Adversarial Transfer Attacks

Yi Cai,Chenyu Wang,Xuefei Ning,Zixuan Zhou,Dimin Niu,Huazhong Yang,Yu Wang

Security is becoming increasingly critical in deep learning applications. Recent researches demonstrate that NN models are vulnerable to adversarial attacks, which can mislead them with only small input perturbations. Moreover, adversaries who know the architecture of victim models can conduct more effective attacks. Unfortunately, the architectural knowledge can usually be stolen by the adversaries by exploiting the system-level hints through many side channels, which is referred to as the neural architecture extraction attack. Conventional countermeasures for neural architecture extraction can introduce large overhead, and different hardware platforms have diverse types of side-channel leakages such that many expert efforts are needed in developing hardware-specific countermeasures. In this paper, we propose DeepGuiser, an automatic, hardware-agnostic, and retrain-free neural architecture disguising method, to disguise the neural architectures to reduce the harm of neural architecture extraction attacks. In a nutshell, given a trained model, DeepGuiser outputs a deploy model that is functionally equivalent with the trained model but with a different (i.e., disguising) architecture. DeepGuiser can minimize the harm of the follow-up adversarial transfer attacks to the deploy model, even if the disguising architecture is completely stolen by the architecture extraction attack. Experiments demonstrate that DeepGuiser can effectively disguise diverse architectures and impede the adversarial transferability by 13.87% ~ 32.59%, while only introducing 10% ~ 40% extra inference latency.

****

FACS: FAST ADAPTIVE CHANNEL SQUEEZING

Ashish Kumar,Laxmidhar Behera

Channel squeezing is one of the central operations performed in CNN bottlenecks to reduce the number of channels in a feature map. This operation is carried out by using a 1 × 1 pointwise convolution which constitutes a significant amount of computations and parameters in a given network. ResNet-50 for instance, consists of 16 such layers which form 33% of total layers and 25% (1.05B/4.12B) of total FLOPs or computations. In the light of their predominance, we propose a novel “Fast Adaptive Channel Squeezing” module which carries out the squeezing operation in a computationally efficient manner. The key benefit of FACS is that it neither alters the number of parameters nor affects the accuracy of a given network. When plugged into diverse CNNs architectures, namely ResNet, VGG, and MobileN et-v2, FACS achieves state-of-the-art performance on ImageNet and CIFAR datasets at dramatically reduced FLOPs. FACS also cuts the training time significantly, and lowers the latency which is particularly advantageous for fast inference on edge devices. The source-code will be made publicly available.

****

Pre-trained Language Models can be Fully Zero-Shot Learners

Xuandong Zhao,Siqi Ouyang,Zhiguo Yu,Ming Wu,Lei Li

How can we extend a pre-trained model to many language understanding tasks, without labeled or additional unlabeled data? Pre-trained language models (PLMs) have been effective for a wide range of NLP tasks. However, existing approaches either require fine-tuning on downstream labeled datasets or manually constructing proper prompts. In this paper, we propose nonparametric prompting PLM (NPPrompt) for fully zero-shot language understanding. Unlike previous methods, NPPrompt uses only pre-trained language models and does not require any labeled data or additional raw corpus for further fine-tuning, nor does it rely on humans to construct a comprehensive set of prompt label words. We evaluate NPPrompt against previous major few-shot and zero-shot learning methods on diverse NLP tasks: including text classification, text entailment, similar text retrieval, and paraphrasing. Experimental results demonstrate that our NPPrompt outperforms the previous best fully zero-shot method by big margins, with absolute gains of 12.8% in accuracy on text classification and 18.9% on the GLUE benchmark.

****

On Compositional Uncertainty Quantification for Seq2seq Graph Parsing

Zi Lin,Du Phan,Panupong Pasupat,Jeremiah Zhe Liu,Jingbo Shang

Recent years have witnessed the success of applying seq2seq models to graph parsing tasks, where the outputs are compositionally structured (e.g., a graph or a tree). However, these seq2seq approaches pose a challenge in quantifying the model’s compositional uncertainty on graph structures due to the gap between seq2seq output probability and structural probability on the graph. This work is the first to quantify and evaluate compositional uncertainty for seq2seq graph parsing tasks. First, we proposed a generic, probabilistically interpretable framework that allows correspondences between seq2seq output probability to structural probability on the graph. This framework serves as a powerful medium for quantifying a seq2seq model's compositional uncertainty on graph elements (i.e., nodes or edges). Second, to evaluate uncertainty quality in terms of calibration, we propose a novel metric called Compositional Expected Calibration Error (CECE) which can measure a model’s calibration behavior in predicting graph structures. By a thorough evaluation for compositional uncertainty on three different tasks across ten domains, we demonstrate that CECE is a better reflection for distributional shift compared to vanilla sequence ECE. Finally, we validate the effectiveness of compositional uncertainty considering the task of collaborative semantic parsing, where the model is allowed to send limited subgraphs for human review. The results show that the collaborative performance based on uncertain subgraph selection consistently outperforms random subgraph selection (30% average error reduction rate) and performs comparably to oracle subgraph selection (only 0.33 difference in average prediction error), indicating that compositional uncertainty is an ideal signal for model errors and can benefit various downstream tasks.

****

Generative Gradual Domain Adaptation with Optimal Transport

Yifei He,Haoxiang Wang,Han Zhao

Unsupervised domain adaptation (UDA) adapts a model from a labeled source domain to an unlabeled target domain in a one-off way. Though widely applied, UDA faces a great challenge whenever the distribution shift between the source and the target is large. Gradual domain adaptation (GDA) mitigates this limitation by using intermediate domains to gradually adapt from the source to the target domain. However, it remains an open problem on how to leverage this paradigm when the oracle intermediate domains are missing or scarce. To approach this practical challenge, we propose Generative Gradual Domain Adaptation with Optimal Transport (GOAT), an algorithmic framework that can generate intermediate domains in a data-dependent way. More concretely, we generate intermediate domains along the Wass erstein geodesic between two given consecutive domains in a feature space, and apply gradual self-training, a standard GDA algorithm, to adapt the source-trained classifier to the target along the sequence of intermediate domains. Empirically, we demonstrate that our GOAT framework can improve the performance of standard GDA when the oracle intermediate domains are scarce, significantly broadening the real-world application scenarios of GDA.

****

Free Lunch for Domain Adversarial Training: Environment Label Smoothing

YiFan Zhang,xue wang,Jian Liang,Zhang Zhang,Liang Wang,Rong Jin,Tieniu Tan

A fundamental challenge for machine learning models is how to generalize learned models for out-of-distribution (OOD) data. Among various approaches, exploiting invariant features by Domain Adversarial Training (DAT) received widespread attention. Despite its success, we observe training instability from DAT, mostly due to over-confident domain discriminator and environment label noise. To address this issue, we proposed Environment Label Smoothing (ELS), which encourages the discriminator to output soft probability, which thus reduces the confidence of the discriminator and alleviates the impact of noisy environment labels. We demonstrate, both experimentally and theoretically, that ELS can improve training stability, local convergence, and robustness to noisy environment labels. By incorporating ELS with DAT methods, we are able to yield state-of-art results on a wide range of domain generalization/adaptation tasks, particularly when the environment labels are highly noisy.

****

Scaling Forward Gradient With Local Losses

Mengye Ren,Simon Kornblith,Renjie Liao,Geoffrey Hinton

Forward gradient learning computes a noisy directional gradient and is a biologically plausible alternative to backprop for learning deep neural networks. The standard forward gradient algorithm suffers from the curse of dimensionality in the number of parameters. In this paper, we propose to scale forward gradient by adding a large number of local greedy loss functions. We consider block-wise, patch-wise, and channel group-wise local losses, and show that activity perturbation reduces variance compared to weight perturbation. Inspired by MLPMixer, we also propose a new architecture, LocalMixer, that is more suitable for local learning. We find local learning can work well with both supervised classification and self-supervised contrastive learning. Empirically, it can match backprop on MNIST and CIFAR-10 and significantly outperform backprop-free algorithms on ImageNet.

****

PAC-NeRF: Physics Augmented Continuum Neural Radiance Fields for Geometry-Agnostic System Identification

Xuan Li,Yi-Ling Qiao,Peter Yichen Chen,Krishna Murthy Jatavallabhula,Ming Lin,Chenfanfu Jiang,Chuang Gan

Existing approaches to system identification (estimating the physical parameters of an object) from videos assume known object geometries. This precludes their applicability in a vast majority of scenes where object geometries are complex or unknown. In this work, we aim to identify parameters characterizing a physical system from a set of multi-view videos without any assumption on object geometry or topology. To this end, we propose "Physics Augmented Continuum Neural Radiance Fields" (PAC-NeRF), to estimate both the unknown geometry and physical parameters of highly dynamic objects from multi-view videos. We design PAC-NeRF to only ever produce physically plausible states by enforcing the neural radiance field to follow the conservation laws of continuum mechanics. For this, we design a hybrid Eulerian-Lagrangian representation of the neural radiance field, i.e., we use the Eulerian grid representation for NeRF density and color fields, while advecting the neural radiance fields via Lagrangian particles. This hybrid Eulerian-Lagrangian representation seamlessly blends efficient neural rendering with the material point method (MPM) for robust differentiable physics simulation. We validate the effectiveness of our proposed framework on geometry and physical parameter estimation over a vast range of materials, including elastic bodies, pl

asticine, sand, Newtonian and non-Newtonian fluids, and demonstrate significant performance gain on most tasks.
****

Linearly Constrained Bilevel Optimization: A Smoothed Implicit Gradient Approach
Prashant Khanduri,Ioannis Tsaknakis,Yihua Zhang,Jia Liu,Sijia Liu,Jiawei Zhang,Mingyi Hong
This work develops analysis and algorithms for solving a class of bilevel optimization problems where the lower-level (LL) problems have linear constraints. Most of the existing approaches for constrained bilevel problems rely on value function based approximate reformulations, which suffer from issues such as non-convex and non-differentiable constraints. In contrast, in this work, we develop an implicit gradient-based approach, which is easy to implement, and is suitable for machine learning applications. We first provide in-depth understanding of the problem, by showing that the implicit objective for such problems is in general non-differentiable. However, if we add some small (linear) perturbation to the LL objective, the resulting problem becomes differentiable almost surely. This key observation opens the door for developing (deterministic and stochastic) gradient-based algorithms similar to the state-of-the-art ones for unconstrained bi-level problems. We show that when the implicit function is assumed to be strongly-convex, convex and non-convex, the resulting algorithms converge with guaranteed rate. Finally, we experimentally corroborate the theoretical findings and evaluate the performance of the proposed framework on numerical and adversarial learning problems. To our knowledge, this is the first time that (implicit) gradient-based methods have been developed and analyzed for the considered class of bilevel problems.
****

Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning
Anton Bakhtin,David J Wu,Adam Lerer,Jonathan Gray,Athul Paul Jacob,Gabriele Farina,Alexander H Miller,Noam Brown
No-press Diplomacy is a complex strategy game involving both cooperation and competition that has served as a benchmark for multi-agent AI research. While self-play reinforcement learning has resulted in numerous successes in purely adversarial games like chess, Go, and poker, self-play alone is insufficient for achieving optimal performance in domains involving cooperation with humans. We address this shortcoming by first introducing a planning algorithm we call DiL-piKL that regularizes a reward-maximizing policy toward a human imitation-learned policy. We prove that this is a no-regret learning algorithm under a modified utility function. We then show that DiL-piKL can be extended into a self-play reinforcement learning algorithm we call RL-DiL-piKL that provides a model of human play while simultaneously training an agent that responds well to this human model. We used RL-DiL-piKL to train an agent we name Diplodocus.
In a 200-game no-press Diplomacy tournament involving 62 human participants spanning skill levels from beginner to expert, two Diplodocus agents both achieved a higher average score than all other participants who played more than two games, and ranked first and third according to an Elo ratings model.
****

Understanding Embodied Reference with Touch-Line Transformer
Yang Li,Xiaoxue Chen,Hao Zhao,Jiangtao Gong,Guyue Zhou,Federico Rossano,Yixin Zhu
We study embodied reference understanding, the task of locating referents using embodied gestural signals and language references. Human studies have revealed that, contrary to popular belief, objects referred to or pointed to do not lie on the elbow-wrist line, but rather on the so-called virtual touch line. Nevertheless, contemporary human pose representations lack the virtual touch line. To tackle this problem, we devise the touch-line Transformer: It takes as input tokenized visual and textual features and simultaneously predicts the referent’s bounding box and a touch-line vector. Leveraging this touch-line prior, we further devise a geometric consistency loss that promotes co-linearity between referents and touch lines. Using the touch line as gestural information dramatically improv

es model performances: Experiments on the YouRefIt dataset demonstrate that our method yields a +25.0% accuracy improvement under the 0.75 IoU criterion, hence closing 63.6% of the performance difference between models and humans. Furthermore, we computationally validate prior human studies by demonstrating that computational models more accurately locate referents when employing the virtual touch line than when using the elbow-wrist line.

****

Evaluating Robustness of Cooperative MARL: A Model-based Approach
Nhan H Pham,Lam M. Nguyen,Jie Chen,Hoang Thanh Lam,Subhro Das,Tsui-Wei Weng

In recent years, a proliferation of methods were developed for cooperative multi-agent reinforcement learning (c-MARL). However, the robustness of c-MARL agents against adversarial attacks has been rarely explored. In this paper, we propose to evaluate the robustness of c-MARL agents via a model-based approach, named c-MBA. Our proposed formulation can craft much stronger adversarial state perturbations of c-MARL agents to lower total team rewards than existing model-free approaches. In addition, we propose the first victim-agent selection strategy and the first data-driven approach to define targeted failure states where each of them allows us to develop even stronger adversarial attack without the expert knowledge to the underlying environment. Our numerical experiments on two representative MARL benchmarks illustrate the advantage of our approach over other baselines: our model-based attack consistently outperforms other baselines in all tested environments.

****

Rethinking Symbolic Regression Datasets and Benchmarks for Scientific Discovery
Yoshitomo Matsubara,Naoya Chiba,Ryo Igarashi,Tatsunori Taniai,Yoshitaka Ushiku

This paper revisits datasets and evaluation criteria for Symbolic Regression, a task of expressing given data using mathematical equations, specifically focused on its potential for scientific discovery. Focused on a set of formulas used in the existing datasets based on Feynman Lectures on Physics, we recreate 120 datasets to discuss the performance of symbolic regression for scientific discovery (SRSD). For each of the 120 SRSD datasets, we carefully review the properties of the formula and its variables to design reasonably realistic sampling range of values so that our new SRSD datasets can be used for evaluating the potential of SRSD such as whether or not an SR method can (re)discover physical laws from such datasets. As an evaluation metric, we also propose to use normalized edit distances between a predicted equation and the ground-truth equation trees. While existing metrics are either binary or errors between the target values and an SR model's predicted values for a given input, normalized edit distances evaluate a sort of similarity between the ground-truth and predicted equation trees. We have conducted experiments on our new SRSD datasets using five state-of-the-art SR methods in SRBench and a simple baseline based on a recent Transformer architecture. The results show that we provide a more realistic performance evaluation and open up a new machine learning-based approach for scientific discovery. We provide our datasets and code as part of the supplementary material.

****

The Cost of Privacy in Fair Machine Learning
Songkai Xue,Yuekai Sun

A common task in fair machine learning is training ML models that preserve certain summary statistics across subpopulations defined by sensitive attributes. However, access to such sensitive attributes in training data is restricted and the learner must rely on noisy proxies for the sensitive attributes. In this paper, we study the effect of a privacy mechanism that obfuscates the sensitive attributes from the learner on the fairness of the resulting classifier. We show that the cost of privacy in fair ML is a decline in the generalizability of fairness constraints.

****

VARIATIONAL ADAPTIVE GRAPH TRANSFORMER FOR MULTIVARIATE TIME SERIES MODELING
Long Tian,Wenchao Chen,Bo Chen,Muyao Wang,Liang Dai,BaoLin Sun,Mingyuan Zhou

Multivariate time series (MTS) are widely collected by large-scale complex systems, such as internet services, IT infrastructures, and wearable devices. The mod eling of MTS has long been an important but challenging task. To capture complex long-range dynamics, Transformers have been utilized in MTS modeling and achieved attractive performance. However, Transformers in general do not well capture the diverse relationships between different channels within MTS and have difficulty in modeling MTS with complex distributions due to the lack of stochasticity. In this paper, we first incorporate relational modeling into Transformer to develop an adaptive Graph Transformer (G-Trans) module for MTS. Then, we further consider stochastity by introducing a powerful embedding guided probabilistic generative module for G-Trans to construct Variational adaptive Graph Transformer (VG-Trans), which is a well-defined variational generative dynamic model. VG-Trans is utilized to learn expressive representations of MTS, being an plug-and-play framework that can be applied to forecasting and anomaly detection tasks of MTS. For efficient inference, we develop an autoencoding variational inference scheme with a combined prediction and reconstruction loss. Extensive experiments on diverse datasets show the efficient of VG-Trans on MTS modeling and improving the existing methods on VG-Trans outperforms state-of-the-art methods on a variety of MTS modeling tasks.

****

Efficient Large-scale Transformer Training via Random and Layerwise Token Dropping

Zhewei Yao,Xiaoxia Wu,Conglong Li,Connor Holmes,Minjia Zhang,Cheng Li,Yuxiong He

Large-scale transformer models have become the de-facto architectures for various machine learning applications, e.g., CV and NLP.

However, those large models also introduce prohibitive training costs.

To mitigate this issue, we propose a novel random and layerwise token dropping method (\OURS), which skips the computation of a subset of the input tokens at all middle layers.

Particularly, \OURS achieves considerable speedups and comparable accuracy as the standard training baseline.

Compared to other token dropping methods, \OURS does not require (1) any importance score-based metrics, (2) any special token treatment (e.g., \texttt{[CLS]}), and (3) many layers in full sequence length training except the first and the last layers.

Besides, a new \layertoken learning rate schedule is proposed for pretraining problems that resolve the heavy tuning requirement for our proposed training mechanism.

Finally, we demonstrate that \OURS can be applied to broader applications, including \gpt and \bert pretraining as well as ViT and \gpt finetuning tasks.

Our results show that \OURS can save about 33.3\% theoretical compute cost and 25.6\% wall-clock training time while achieving similar zero-shot evaluations on \gptb as compared to baseline.

****

Demystifying black-box DNN training processes through Concept-Monitor

Mohammad Ali Khan,Tuomas Oikarinen,Tsui-Wei Weng

Despite the successes of deep neural networks (DNNs) on a broad range of tasks little has been understood of why and how they achieve such victories due to their complex architecture and their opaque black-box training processes. With the goal to unveil the mystery of DNNs, in this work, we propose a general framework called Concept-Monitor to uncover the black-box DNN training processes automatically for the first time. Our proposed Concept-Monitor enables human-interpretable visualization on the DNN training processes and thus facilitates transparency as well as deeper understanding of how DNNs function and operate along the training iterations. Using Concept-Monitor, we are able to observe and compare different training paradigms at ease, including standard training, fine-tuning, adversarial training and network pruning for Lottery Ticket Hypothesis, which brings new insights on why and how adversarial training and network pruning work and how they modify the network during training. For example, we find that the lottery ticket hypothesis discovers a mask that makes neurons interpretable at initialization, \textit{without} any finetuning, and we also found that adversarially robust models have more neurons relying on color as compared to standard models tra ined on the same dataset.

****

Large Language Models Can Self-improve

Jiaxin Huang,Shixiang Shane Gu,Le Hou,Yuexin Wu,Xuezhi Wang,Hongkun Yu,Jiawei Han

Large Language Models (LLMs) have achieved excellent performances in various tasks. However, fine-tuning an LLM requires extensive supervision. Human, on the other hand, may improve their reasoning abilities by self-thinking without external inputs. In this work, we demonstrate that an LLM is also capable of self-improving with only unlabeled datasets. We use a pre-trained LLM to generate “high-confidence” rationale-augmented answers for unlabeled questions using Chain-of-Thought prompting and self-consistency, and fine-tune the LLM using those self-generated solutions as target outputs. We show that our approach improves the general reasoning ability of a 540B-parameter LLM (74.4%→82.1% on GSM8K, 78.2%→83.0% on DROP, 90.0%→94.4% on OpenBookQA, and 63.4%→67.9% on ANLI-A3) and achieves state-of-the-art-level performance, without any ground truth label. We conduct ablation studies and show that finetuning on reasoning is critical for self-improvement.

****

Calibration Matters: Tackling Maximization Bias in Large-scale Advertising Recommendation Systems

Yewen Fan,Nian Si,Kun Zhang

Calibration is defined as the ratio of the average predicted click rate to the true click rate. The optimization of calibration is essential to many online advertising recommendation systems because it directly affects the downstream bids in ads auctions and the amount of money charged to advertisers. Despite its importance, calibration often suffers from a problem called “maximization bias”. Maximization bias refers to the phenomenon that the maximum of predicted values overestimates the true maximum. The problem is introduced because the calibration is computed on the set selected by the prediction model itself. It persists even if unbiased predictions are achieved on every datapoint and worsens when covariate shifts exist between the training and test sets. To mitigate this problem, we quantify maximization bias and propose a variance-adjusting debiasing (VAD) meta-algorithm in this paper. The algorithm is efficient, robust, and practical as it is able to mitigate maximization bias problem under covariate shifts, without incurring additional online serving costs or compromising the ranking performance. We demonstrate the effectiveness of the proposed algorithm using a state-of-the-art recommendation neural network model on a large-scale real-world dataset.

****

Excess risk analysis for epistemic uncertainty with application to variational inference

Futoshi Futami,Tomoharu Iwata,Naonori Ueda,Issei Sato,Masashi Sugiyama

Bayesian deep learning plays an important role especially for its ability evaluating epistemic uncertainty (EU). Due to computational complexity issues, approximation methods such as variational inference (VI) have been used in practice to obtain posterior distributions and their generalization abilities have been analyzed extensively, for example, by PAC-Bayesian theory; however, little analysis exists on EU, although many numerical experiments have been conducted on it.

In this study, we analyze the EU of supervised learning in approximate Bayesian inference by focusing on its excess risk. First, we theoretically show the novel relations between generalization error and the widely used EU measurements, such as the variance and mutual information of predictive distribution, and derive their convergence behaviors. Next, we clarify how the objective function of VI regularizes the EU. With this analysis, we propose a new objective function for VI that directly controls the prediction performance and the EU based on the PAC-Bayesian theory. Numerical experiments show that our algorithm significantly improves the EU evaluation over the existing VI methods.

****

Memorization-Dilation: Modeling Neural Collapse Under Noise

Duc Anh Nguyen,Ron Levie,Julian Lienen,Eyke Hüllermeier,Gitta Kutyniok

The notion of neural collapse refers to several emergent phenomena that have been empirically observed across various canonical classification problems. During the terminal phase of training a deep neural network, the feature embedding of all examples of the same class tend to collapse to a single representation, and the features of different classes tend to separate as much as possible. Neural collapse is often studied through a simplified model, called the layer-peeled model, in which the network is assumed to have ``infinite expressivity'' and can map each data point to any arbitrary representation. In this work we study a more realistic variant of the layer-peeled model, which takes the positivity of the features into account. Furthermore, we extend this model to also incorporate the limited expressivity of the network. Empirical evidence suggests that the memorization of noisy data points leads to a degradation (dilation) of the neural collapse. Using a model of the memorization-dilation (M-D) phenomenon, we show one mechanism by which different losses lead to different performances of the trained network on noisy data. Our proofs reveal why label smoothing, a modification of cross-entropy empirically observed to produce a regularization effect, leads to improved generalization in classification tasks.

****

Spacetime Representation Learning

Marc T. Law,James Lucas

Much of the data we encounter in the real world can be represented as directed graphs. In this work, we introduce a general family of representations for directed graphs through connected time-oriented Lorentz manifolds, called "spacetimes" in general relativity. Spacetimes intrinsically contain a causal structure that indicates whether or not there exists a causal or even chronological order between points of the manifold, called events. This chronological order allows us to naturally represent directed edges via imposing the correct ordering when the nodes are embedded as events in the spacetime. Previous work in machine learning only considers embeddings lying on the simplest Lorentz manifold or does not exploit the connection between Lorentzian pre-length spaces and directed graphs. We introduce a well-defined approach to map data onto a general family of spacetimes. We empirically evaluate our framework in the tasks of hierarchy extraction of undirected graphs, directed link prediction and representation of directed graphs.

****

Meta-Learning General-Purpose Learning Algorithms with Transformers

Louis Kirsch,James Harrison,Jascha Sohl-Dickstein,Luke Metz

Modern machine learning requires system designers to specify aspects of the learning pipeline, such as losses, architectures, and optimizers. Meta-learning, or learning-to-learn, instead aims to learn those aspects, and promises to unlock greater capabilities with less manual effort. One particularly ambitious goal of meta-learning is to train general purpose learning algorithms from scratch, using only black box models with minimal inductive bias. A general purpose learning algorithm is one which takes in training data, and produces test-set predictions across a wide range of problems, without any explicit definition of an inference model, training loss, or optimization algorithm. In this paper we show that Transformers and other black-box models can be meta-trained to act as general purpose learning algorithms, and can generalize to learn on different datasets than used during meta-training. We characterize phase transitions between algorithms that generalize, algorithms that memorize, and algorithms that fail to meta-train at all, induced by changes in model size, number of tasks used during meta-training, and meta-optimization hyper-parameters. We further show that the capabilities of meta-trained algorithms are bottlenecked by the accessible state size determining the next prediction, unlike standard models which are thought to be bottlenecked by parameter count. Finally, we propose practical interventions such as biasing the training distribution that improve the meta-training and meta-generalization of general purpose learning algorithms.

****

Learning to Extrapolate: A Transductive Approach

Aviv Netanyahu,Abhishek Gupta,Max Simchowitz,Kaiqing Zhang,Pulkit Agrawal

Machine learning systems, especially with overparameterized deep neural networks, can generalize to novel test instances drawn from the same distribution as the training data. However, they fare poorly when evaluated on \emph{out-of-support} test points. In this work, we tackle the problem of developing machine learning systems that retain the power of overparameterized function approximators while enabling extrapolation to out-of-support test points when possible. This is accomplished by noting that under certain conditions, a ``transductive'' reparameterization can convert an out-of-support extrapolation problem into a problem of within-support combinatorial generalization. We propose a simple strategy based on bilinear embeddings to enable this type of combinatorial generalization, thereby addressing the out-of-support extrapolation problem under certain conditions. We instantiate a simple, practical algorithm applicable to various supervised learning and imitation learning tasks.
****

Label-free Concept Bottleneck Models

Tuomas Oikarinen,Subhro Das,Lam M. Nguyen,Tsui-Wei Weng

Concept bottleneck models (CBM) are a popular way of creating more interpretable neural networks by having hidden layer neurons correspond to human-understandable concepts. However, existing CBMs and their variants have two crucial limitations: first, they need to collect labeled data for each of the predefined concepts, which is time consuming and labor intensive; second, the accuracy of a CBM is often significantly lower than that of a standard neural network, especially on more complex datasets. This poor performance creates a barrier for adopting CBMs in practical real world applications. Motivated by these challenges, we propose Label-free CBM which is a novel framework to transform any neural network into an interpretable CBM without labeled concept data, while retaining a high accuracy. Our Label-free CBM has many advantages, it is: scalable - we present the first CBM scaled to ImageNet, efficient - creating a CBM takes only a few hours even for very large datasets, and automated - training it for a new dataset requires minimal human effort. Our code is available at https://github.com/Trustworthy-ML-Lab/Label-free-CBM.
****

Multi-level Protein Structure Pre-training via Prompt Learning

Zeyuan Wang,Qiang Zhang,Shuang-Wei HU,Haoran Yu,Xurui Jin,Zhichen Gong,Huajun Chen

A protein can focus on different structure levels to implement its functions. Each structure has its own merit and driving forces in describing some specific characteristics, and they cannot replace each other. Most existing function prediction methods take the tertiary structure as input, unintentionally ignoring the other levels of protein structures. Considering protein sequences can determine multi-level structures, in this paper, we aim to realize the comprehensive potential of protein sequences for function prediction. Specifically, we propose a new prompt-guided multi-task pre-training and fine-tuning framework, and the resulting protein model is called PromptProtein. Through the prompt-guided multi-task pre-training, we learn multiple prompt signals to steer the model to focus on different structure levels. We also design a prompt fine-tuning module to provide downstream tasks the on-demand flexibility of utilizing respective levels of structure information. Extensive experiments on function prediction and protein engineering show that PromptProtein outperforms state-of-the-art methods by large margins.
****

CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks

Tuomas Oikarinen,Tsui-Wei Weng

In this paper, we propose CLIP-Dissect, a new technique to automatically describe the function of individual hidden neurons inside vision networks. CLIP-Dissect leverages recent advances in multimodal vision/language models to label internal neurons with open-ended concepts without the need for any labeled data or human examples. We show that CLIP-Dissect provides more accurate descriptions than existing methods for last layer neurons where the ground-truth is available as we ll as qualitatively good descriptions for hidden layer neurons. In addition, our method is very flexible: it is model agnostic, can easily handle new concepts and can be extended to take advantage of better multimodal models in the future. Finally CLIP-Dissect is computationally efficient and can label all neurons from five layers of ResNet-50 in just 4 minutes, which is more than 10$\times$ faster than existing methods. Our code is available at https://github.com/Trustworthy-ML-Lab/CLIP-dissect.

****

GLM-130B: An Open Bilingual Pre-trained Model

Aohan Zeng,Xiao Liu,Zhengxiao Du,Zihan Wang,Hanyu Lai,Ming Ding,Zhuoyi Yang,Yifan Xu,Wendi Zheng,Xiao Xia,Weng Lam Tam,Zixuan Ma,Yufei Xue,Jidong Zhai,Wenguang Chen,Zhiyuan Liu,Peng Zhang,Yuxiao Dong,Jie Tang

We introduce GLM-130B, a bilingual (English and Chinese) pre-trained language model with 130 billion parameters. It is an attempt to open-source a 100B-scale model as good as GPT-3 (davinci) and unveil how models of such a scale can be successfully pre-trained. Over the course of this effort, we face numerous unexpected technical and engineering challenges, particularly on loss spikes and divergence. In this paper, we introduce the pre-training process of GLM-130B including its design choices, training strategies for both efficiency and stability, and engineering efforts. The resultant GLM-130B model offers significant outperformance over GPT-3 175B on a wide range of popular English benchmarks while the performance advantage is not observed in OPT-175B and BLOOM-176B. It also consistently and significantly outperforms ERNIE TITAN 3.0 260B—the largest Chinese language model—across related benchmarks. Finally, we leverage a unique scaling property of GLM-130B to reach INT4 quantization with almost no performance loss, making it the first among 100B-scale models and more importantly, allowing its effective inference on 4×RTX 3090 (24G) or 8×RTX 2080 Ti (11G) GPUs, the most ever affordable GPUs required for using 100B-scale models. The GLM-130B model weights are publicly accessible and its code, training logs, related toolkit, and lessons learned are open-sourced at https://github.com/THUDM/GLM-130B/.

****

Causal Estimation for Text Data with (Apparent) Overlap Violations

Lin Gui,Victor Veitch

Consider the problem of estimating the causal effect of some attribute of a text document; for example: what effect does writing a polite vs. rude email have on response time? To estimate a causal effect from observational data, we need to adjust for confounding aspects of the text that affect both the treatment and outcome---e.g., the topic or writing level of the text. These confounding aspects are unknown a priori, so it seems natural to adjust for the entirety of the text (e.g., using a transformer). However, causal identification and estimation procedures rely on the assumption of overlap: for all levels of the adjustment variables, there is randomness leftover so that every unit could have (not) received treatment. Since the treatment here is itself an attribute of the text, it is perfectly determined, and overlap is apparently violated. The purpose of this paper is to show how to handle causal identification and obtain robust causal estimation in the presence of apparent overlap violations. In brief, the idea is to use supervised representation learning to produce a data representation that preserves confounding information while eliminating information that is only predictive of the treatment. This representation then suffices for adjustment and satisfies overlap. Adapting results on non-parametric estimation, we show that this procedure shows robustness with respect to conditional outcome misestimation and yields a low-bias estimator that admits valid uncertainty quantification under weak conditions. Empirical results show reductions in bias and strong improvements in uncertainty quantification relative to the natural (transformer-based) baseline.

****

Understanding Pruning at Initialization: An Effective Node-Path Balancing Perspective

Hoang Pham,Anh Ta,Shiwei Liu,Dung D. Le,Long Tran-Thanh

Pruning at initialization (PaI) methods aim to remove weights of neural networks before training in pursuit of reducing training costs. While current PaI methods are promising and outperform random pruning, much work remains to be done to understand and improve PaI methods to achieve the performance of pruning after training. In particular, recent studies (Frankle et al., 2021; Su et al., 2020) present empirical evidence for the potential of PaI, and show intriguing properties like layerwise random shuffling connections of pruned networks preserves or even improves the performance. Our paper gives new perspectives on PaI from the geometry of subnetwork configurations. We propose to use two quantities to probe the shape of subnetworks: the numbers of effective paths and effective nodes (or channels). Using these numbers, we provide a principled framework to better understand PaI methods. Our main findings are: (i) the width of subnetworks matters in regular sparsity levels (< 99%) - this matches the competitive performance of shuffled layerwise subnetworks; (ii) node-path balancing plays a critical role in the quality of PaI subnetworks, especially in extreme sparsity regimes. These innovate an important direction to network pruning that takes into account the subnetwork topology itself. To illustrate the promise of this direction, we present a fairly naive method based on SynFlow (Tanaka et al., 2020) and conduct extensive experiments on different architectures and datasets to demonstrate its effectiveness.

****

Data Continuity Matters: Improving Sequence Modeling with Lipschitz Regularizer

Eric Qu,Xufang Luo,Dongsheng Li

Sequence modeling is a core problem in machine learning, and various neural networks have been designed to process different types of sequence data. However, few attempts have been made to understand the inherent data property of sequence data, neglecting the critical factor that may significantly affect the performance of sequence modeling. In this paper, we theoretically and empirically analyze a generic property of sequence data, i.e., continuity, and connect this property with the performance of deep models. First, we empirically observe that different kinds of models for sequence modeling prefer data with different continuity. Then, we theoretically analyze the continuity preference of different models in both time and frequency domains. To further utilize continuity to improve sequence modeling, we propose a simple yet effective Lipschitz Regularizer, that can flexibly adjust data continuity according to model preferences, and bring very little extra computational cost. Extensive experiments on various tasks demonstrate that altering data continuity via Lipschitz Regularizer can largely improve the performance of many deep models for sequence modeling.

****

MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations

Nicklas Hansen,Yixin Lin,Hao Su,Xiaolong Wang,Vikash Kumar,Aravind Rajeswaran

Poor sample efficiency continues to be the primary challenge for deployment of deep Reinforcement Learning (RL) algorithms for real-world applications, and in particular for visuo-motor control. Model-based RL has the potential to be highly sample efficient by concurrently learning a world model and using synthetic rollouts for planning and policy improvement. However, in practice, sample-efficient learning with model-based RL is bottlenecked by the exploration challenge. In this work, we find that leveraging just a handful of demonstrations can dramatically improve the sample-efficiency of model-based RL. Simply appending demonstrations to the interaction dataset, however, does not suffice. We identify key ingredients for leveraging demonstrations in model learning -- policy pretraining, targeted exploration, and oversampling of demonstration data -- which forms the three phases of our model-based RL framework. We empirically study three complex visuo-motor control domains and find that our method is 160%-250% more successful in completing sparse reward tasks compared to prior approaches in the low data regime (100k interaction steps, 5 demonstrations). Code and videos are available at https://nicklashansen.github.io/modemrl.

****

Improving the Estimation of Instance-dependent Transition Matrix by using Self-supervised Learning

Yexiong Lin,Yu Yao,Zhaoqing Wang,Xu Shen,Jun Yu,Bo Han,Tongliang Liu

The transition matrix reveals the transition relationship between clean labels and noisy labels. It plays an important role in building statistically consistent classifiers. In real-world applications, the transition matrix is usually unknown and has to be estimated. It is a challenging task to accurately estimate the transition matrix, especially when it depends on the instance. Given that both instances and noisy labels are available, the major difficulty of learning the transition matrix comes from the missing of clean information. A lot of methods have been proposed to infer clean information. The self-supervised learning has demonstrated great success. These methods could even achieve comparable performance with supervised learning on some datasets but without requiring any labels during the training. It implies that these methods can efficiently infer clean labels.

Motivated by this, in this paper, we have proposed a practical method that leverages self-supervised learning to help learn the instance-dependent transition matrix. Empirically, the proposed method has achieved state-of-the-art performance on different datasets.

****

Holographic-(V)AE: an end-to-end SO(3)-Equivariant (Variational) Autoencoder in Fourier Space

Gian Marco Visani,Michael Neal Pun,Armita Nourmohammad

Group-equivariant neural networks have emerged as a data-efficient approach to solve classification and regression tasks, while respecting the relevant symmetries of the data. However, little work has been done to extend this paradigm to the unsupervised and generative domains. Here, we present Holographic-(V)AE (H-(V)AE), a fully end-to-end SO(3)-equivariant (variational) autoencoder in Fourier space, suitable for unsupervised learning and generation of data distributed around a specified origin. H-(V)AE is trained to reconstruct the spherical Fourier encoding of data, learning in the process a latent space with a maximally informative invariant embedding alongside an equivariant frame describing the orientation of the data. We extensively test the performance of H-(V)AE on diverse datasets and show that its latent space efficiently encodes the categorical features of spherical images and structural features of protein atomic environments. Our work can further be seen as a case study for equivariant modeling of a data distribution by reconstructing its Fourier encoding.

****

A general differentially private learning framework for decentralized data

Mei Li,Bo Pan,Yafei Wang,Lingchen Kong,Linglong Kong,Bei Jiang

Decentralized consensus learning has been hugely successful, which minimizes a finite sum of expected objective functions over a network of nodes. However, the local communication across neighboring nodes in the network may lead to the leakage of private information. To address this challenge, we propose a general differentially private (DP) learning framework for decentralized data that applies to many non-smooth learning problems. We show that the proposed algorithm retains the performance guarantee in terms of stability, generalization, and finite sample performance. We investigate the impact of local privacy-preserving computation on the global DP guarantee. Further, we extend the discussion by adopting a new class of noise-adding DP mechanisms based on generalized Gaussian distributions to improve the utility-privacy trade-offs. Our numerical results demonstrate the effectiveness of our algorithm and its better performance over the state-of-the-art baseline methods in various decentralized settings.

****

Wasserstein Barycenter-based Model Fusion and Linear Mode Connectivity of Neural Networks

Aditya Kumar Akash,Sixu Li,Nicolas Garcia Trillos

Based on the concepts of Wasserstein barycenter (WB) and Gromov-Wasserstein barycenter (GWB), we propose a unified mathematical framework for neural network (NN) model fusion and utilize it to reveal new insights about the linear mode connectivity of SGD solutions. In our framework, the fusion occurs in a layer-wise manner and builds on an interpretation of a node in a network as a function of the layer preceding it.

The versatility of our mathematical framework allows us to talk about model fusion and linear mode connectivity for a broad class of NNs, including fully connected NN, CNN, ResNet, RNN, and LSTM, in each case exploiting the specific structure of the network architecture. We present extensive numerical experiments to: 1) illustrate the strengths of our approach in relation to other model fusion methodologies and 2) from a certain perspective, provide new empirical evidence for recent conjectures which say that two local minima found by gradient-based methods end up lying on the same basin of the loss landscape after a proper permutation of weights is applied to one of the models.

****

Weakly-Supervised Domain Adaptation in Federated Learning

Enyi Jiang,Oluwasanmi O Koyejo

Federated domain adaptation (FDA) describes the setting where a set of source clients seek to optimize the performance of a target client. To be effective, FDA must address some of the distributional challenges of Federated learning (FL). For instance, FL systems exhibit distribution shifts across clients. Further, labeled data are not always available among the clients. To this end, we propose and compare novel approaches for FDA, combining the few labeled target samples with the source data when auxiliary labels are available to the clients. The in-distribution auxiliary information is included during local training to boost out-of-domain accuracy. Also, during fine-tuning, we devise a simple yet efficient gradient projection method to detect the valuable components from each source client model towards the target direction. The extensive experiments on medical imaging datasets show that our proposed framework significantly improves federated domain adaptation performance.

****

PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm

Toygun Basaklar,Suat Gumussoy,Umit Ogras

Multi-objective reinforcement learning (MORL) approaches have emerged to tackle many real-world problems with multiple conflicting objectives by maximizing a joint objective function weighted by a preference vector. These approaches find fixed customized policies corresponding to preference vectors specified during training. However, the design constraints and objectives typically change dynamically in real-life scenarios. Furthermore, storing a policy for each potential preference is not scalable. Hence, obtaining a set of Pareto front solutions for the entire preference space in a given domain with a single training is critical. To this end, we propose a novel MORL algorithm that trains a single universal network to cover the entire preference space scalable to continuous robotic tasks. The proposed approach, Preference-Driven MORL (PD-MORL), utilizes the preferences as guidance to update the network parameters. It also employs a novel parallelization approach to increase sample efficiency. We show that PD-MORL achieves up to 25% larger hypervolume for challenging continuous control tasks and uses an order of magnitude fewer trainable parameters compared to prior approaches.

****

When Majorities Prevent Learning: Eliminating Bias to Improve Worst-group and Out-of-distribution Generalization

Yu Yang,Gintare Karolina Dziugaite,Baharan Mirzasoleiman

Modern neural networks trained on large datasets have achieved state-of-the-art (in-distribution) generalization performance on various tasks. However, their good generalization performance has been shown to be contributed largely to overfitting spurious biases in large datasets. This is evident by the poor generalization performance of such models on minorities and out-of-distribution data. To alleviate this issue, subsampling the majority groups has been shown to be very effective. However, it is not clear how to find the subgroups (e.g. within a class) in large real-world datasets. Besides, naively subsampling the majority groups can entirely deplete some of their smaller sub-populations and drastically harm the in-distribution performance. Here, we show that tracking gradient trajectories of examples in initial epochs allows for finding large subpopulations of data points. We leverage this observation and propose an importance sampling method that is biased towards selecting smaller subpopulations, and eliminates bias in the large subpopulations. Our experiments confirm the effectiveness of our approach in eliminating spurious biases and learning higher-quality models with superior in- and out-of-distribution performance on various datasets.

****

Understanding the Role of Nonlinearity in Training Dynamics of Contrastive Learning

Yuandong Tian

While the empirical success of self-supervised learning (SSL) heavily relies on the usage of deep nonlinear models, existing theoretical works on SSL understanding still focus on linear ones. In this paper, we study the role of nonlinearity in the training dynamics of contrastive learning (CL) on one and two-layer nonlinear networks with homogeneous activation $h(x) = h'(x)x$. We have two major theoretical discoveries. First, the presence of nonlinearity can lead to many local optima even in 1-layer setting, each corresponding to certain patterns from the data distribution, while with linear activation, only one major pattern can be learned. This suggests that models with lots of parameters can be regarded as a \emph{brute-force} way to find these local optima induced by nonlinearity. Second, in the 2-layer case, linear activation is proven not capable of learning specialized weights into diverse patterns, demonstrating the importance of nonlinearity. In addition, for 2-layer setting, we also discover \emph{global modulation}: those local patterns discriminative from the perspective of global-level patterns are prioritized to learn, further characterizing the learning process. Simulation verifies our theoretical findings.

****

Oracle-oriented Robustness: Robust Image Model Evaluation with Pretrained Models as Surrogate Oracle

Peiyan Zhang,Sunghun Kim,Eric Xing,Haohan Wang

Machine learning has demonstrated remarkable performances over finite datasets, yet whether the scores over the fixed benchmarks can sufficiently indicate the model’s performances 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. 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 causal structure the original test image represents, constrained by a surrogate oracle model pretrained with a large amount of samples. As a result, our new method will offer us a new way to evaluate the models’ robustness performances, 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.

****

Certified Robustness on Structural Graph Matching

Huaqing Shao,Lanjun Wang,Yongwei Wang,Qibing Ren,Junchi Yan

The vulnerability of graph matching (GM) to adversarial attacks has received increasing attention from emerging empirical studies, while the certified robustness of GM has not been explored. Motivated by randomized smoothing, we are the first to define certified robustness on GM and design a new certification strategy called Structure-based Certified Robustness of Graph Matching (SCR-GM). Structural prior information of nodes is used to construct a joint smoothing distribution matrix with physical significance, which certifies a wider range than those obtained by previous iterative optimization methods. Furthermore, we propose a certified space that can be used to derive a strictly certified radius and two radii for evaluation. Experimental results on graph matching datasets reveal that our strategy achieves state-of-the-art ℓ_{2} certified accuracy and regions.

****

CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis

Erik Nijkamp,Bo Pang,Hiroaki Hayashi,Lifu Tu,Huan Wang,Yingbo Zhou,Silvio Savarese,Caiming Xiong

Program synthesis strives to generate a computer program as a solution to a given problem specification, expressed with input-output examples or natural language descriptions. The prevalence of large language models advances the state-of-the-art for program synthesis, though limited training resources and data impede open access to such models. To democratize this, we train and release a family of large language models up to 16.1B parameters, called CODEGEN, on natural language and programming language data, and open source the training library JAXFORMER. We show the utility of the trained model by demonstrating that it is competitive with the previous state-of-the-art on zero-shot Python code generation on HumanEval. We further investigate the multi-step paradigm for program synthesis, where a single program is factorized into multiple prompts specifying subproblems. To this end, we construct an open benchmark, Multi-Turn Programming Benchmark (MTPB), consisting of 115 diverse problem sets that are factorized into multi-turn prompts. Our analysis on MTPB shows that the same intent provided to CODEGEN in multi-turn fashion significantly improves program synthesis over that provided as a single turn. We make the training library JAXFORMER and model checkpoints available as open source contribution: https://github.com/salesforce/CodeGen.

****

Bayesian Optimal Experimental Design for the Survey Bandit Setting

Sang T. Truong,Willie Neiswanger,Susan Athey

The contextual bandit is a classic problem in sequential decision making under uncertainty that finds broad application to tasks in precision medicine, personalized education, and drug discovery. Here, a decision maker repeatedly receives a context, takes an action, and then observes an associated outcome, with the goal of choosing actions that achieve a minimal regret. However, in many settings, the context is not given, and the decision maker must instead collect some information to infer a context before proceeding. For example, when a doctor does not have prior information about a patient, they might ask a sequence of questions before recommending a medical treatment. In this paper, we aim to develop methods for this setting—which we refer to as the \emph{survey bandit}—where the decision maker is not given access to the context but can ask a finite sequence of questions to gain information about the context before taking an action and observing an outcome. Using insights from Bayesian optimal experimental design (BOED) and decision-theoretic information theory, we view the interaction with each user as a BOED task, where the goal is to ask a sequence of questions that elicit the most information about the optimal action for this user. Our procedure is agnostic to the choice of probabilistic model, and we demonstrate its usefulness in a few common classes of distributions. Our algorithm achieves significantly better performance on both synthetic and real data relative to existing baseline methods while remaining statistically efficient, interpretable, and computationally friendly.

****

Synchronized Contrastive Pruning for Efficient Self-Supervised Learning

Jian Meng,Li Yang,Deliang Fan,Jinwoo Shin,Jae-sun Seo

Various self-supervised learning (SSL) methods have demonstrated strong capability in learning visual representations from unlabeled data. However, the current state-of-the-art (SoTA) SSL methods largely rely on heavy encoders to achieve comparable performance as the supervised learning counterpart. Despite the well-learned visual representations, the large-sized encoders impede the energy efficient computation, especially for resource-constrained edge computing. Prior works have utilized the sparsity-induced asymmetry to enhance the contrastive learning of dense models, but the generality between asymmetric sparsity and self-supervised learning has not been fully discovered. Furthermore, transferring the supervised sparse learning to SSL is also largely under-explored. To address the research gap in prior works, this paper investigates the correlation between in-training sparsity and SSL. In particular, we propose a novel sparse SSL algorithm, e mbracing the benefits of contrastiveness while exploiting high sparsity during SSL training. The proposed framework is evaluated comprehensively with various granularities of sparsity, including element-wise sparsity, GPU-friendly N:M structured fine-grained sparsity, and hardware-specific structured sparsity. Extensive experiments across multiple datasets are performed, where the proposed method shows superior performance against the SoTA sparse learning algorithms with various SSL frameworks. Furthermore, the training speedup aided by the proposed method is evaluated with an actual DNN training accelerator model.

****

VEHICLE-INFRASTRUCTURE COOPERATIVE 3D DETECTION VIA FEATURE FLOW PREDICTION

Haibao Yu,Yingjuan Tang,Jilei Mao,Enze Xie,Jirui Yuan,Ping Luo,Zaiqing Nie

Effectively utilizing data from infrastructure could greatly improve autonomous driving safety. Vehicle-Infrastructure Cooperative 3D Object Detection (VIC3D) is an important task to localize and recognize objects surrounding the ego-vehicle by combining the sensor data from both ego-vehicle and roadside infrastructure. However, there are serious temporal asynchrony problems between vehicle and infrastructure data. To the best of our knowledge, no existing work in the literature could effectively solve the asynchrony problem with limited communication bandwidth and computational resources on vehicle-infrastructure devices. This work proposes a novel approach for VIC3D, called Feature Flow Network(FFNet), to effectively address the problem of temporal asynchrony caused by different sensor initialization and latency. Compared with previous feature fusion approaches that only use the current static feature, FFNet transmits the feature flow and generates the future features on-the-fly, aligned with the ego-vehicle timestamp. We propose a self-supervised method to train the feature flow generation model, and use the pre-trained infrastructure model to extract features from randomly assigned future frames as ground truth. Extensive experiments on the DAIR-V2X dataset (a large-scale real-world V2X dataset) show that FFNet establishes a new state of the art, surpassing SOTA methods by up to 5% mAP while with comparable transmission cost. In particularly, FFNet can even make up for almost all the performance drop caused by the temporal asynchrony in 200ms delay.

****

M-L2O: Towards Generalizable Learning-to-Optimize by Test-Time Fast Self-Adaptation

Junjie Yang,Xuxi Chen,Tianlong Chen,Zhangyang Wang,Yingbin Liang

Learning to Optimize (L2O) has drawn increasing attention as it often remarkably accelerates the optimization procedure of complex tasks by "overfitting" specific task type, leading to enhanced performance compared to analytical optimizers. Generally, L2O develops a parameterized optimization method (i.e., "optimizer") by learning from solving sample problems. This data-driven procedure yields L2O that can efficiently solve problems similar to those seen in training, that is, drawn from the same "task distribution". However, such learned optimizers often struggle when new test problems come with a substantially deviation from the training task distribution. This paper investigates a potential solution to this open challenge, by meta-training an L2O optimizer that can perform fast test-time self-adaptation to a out-of-distribution task, in only a few steps. We theoretically characterize the generalization of L2O, and further show that our proposed framework (termed as M-L2O) provably facilitates rapid task adaptation by locating well-adapted initial points for the optimizer weight. Empirical observations on several classic tasks like LASSO and Quadratic, demonstrate that M-L2O converges significantly faster than vanilla L2O with only 5 steps of adaptation, echoing our theoretical results. Codes are available in https://github.com/VITA-Group/M-L2O.

****

ReG-NAS: Graph Neural Network Architecture Search using Regression Proxy Task

Boyi Wei,Cong Hao

Neural Architecture Search (NAS) has become a focus that has been extensively researched in recent years. Innovative achievements are yielded from the area like convolutional neural networks (CNN), recurrent neural networks (RNN) and so on. However, research on NAS for graph neural networks (GNN) is still in a prelimin ary stage. Because of the special structure of graph data, some conclusions drew from CNN cannot be directly applied to GNN. At the same time, for NAS, the models' ranking stability is of great importance for it reflects the reliability of the NAS performance. Unfortunately, little research attention has been paid to it, making it a pitfall in the development of NAS research. In this paper, we proposed a novel NAS pipeline, ReG-NAS, which balances stability, reliability and time cost to search the best GNN architecture. Besides, for the first time, we systematically analyzed factors that will affect models' ranking stability in a given search space, which can be used as a guideline for subsequent studies. Our codes are available at https://anonymous.4open.science/r/ReG-NAS-4D21

****

Mesh-Independent Operator Learning for PDEs using Set Representations

Seungjun Lee

Operator learning, learning the mapping between infinite-dimensional function spaces, has been attracted as an alternative approach to traditional numerical methods to solve partial differential equations (PDEs). In practice, the functions of the physical systems are often observed by sparse or even irregularly distributed measurements, thus the functions are discretized and usually represented by finite structured arrays, which are given as data of input-output pairs. Through training with the arrays, the solution of the trained models should be independent of the discretization of the input function and can be queried at any point continuously. Therefore, the architectures for operator learning should be flexibly compatible with arbitrary sizes and locations of the measurements, otherwise, it can restrict the scalability when the observations have discrepancies between measurement formats. In this paper, we propose to treat the discretized functions as set-valued data and construct an attention-based model, called mesh-independent operator learner (MIOL), to provide proper treatments of input functions and query coordinates for the solution functions by detaching the dependencies on input and output meshes. Our models pre-trained with benchmark datasets of operator learning are evaluated by downstream tasks to demonstrate the generalization abilities to varying discretization formats of the system, which are natural characteristics of the continuous solution of the PDEs.

****

ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning

Olga Golovneva,Moya Peng Chen,Spencer Poff,Martin Corredor,Luke Zettlemoyer,Maryam Fazel-Zarandi,Asli Celikyilmaz

Large language models show improved downstream task performance when prompted to generate step-by-step reasoning to justify their final answers. These reasoning steps greatly improve model interpretability and verification, but objectively studying their correctness (independent of the final answer) is difficult without reliable methods for automatic evaluation. We simply do not know how often the stated reasoning steps actually support the final end task predictions. In this work, we present ROSCOE, a suite of interpretable, unsupervised automatic scores that improve and extend previous text generation evaluation metrics. To evaluate ROSCOE against baseline metrics, we design a typology of reasoning errors and collect synthetic and human evaluation scores on commonly used reasoning datasets. In contrast with existing metrics, ROSCOE can measure semantic consistency, logicality, informativeness, fluency, and factuality — among other traits — by leveraging properties of step-by-step rationales. We empirically verify the strength of our metrics on five human annotated and six programmatically perturbed diagnostics datasets - covering a diverse set of tasks that require reasoning skills and show that ROSCOE can consistently outperform baseline metrics.

****

Robust Multi-Agent Reinforcement Learning against Adversaries on Observation

Chenghe Wang,Yuhang Ran,Lei Yuan,Yang Yu,Zongzhang Zhang

With the broad applications of deep learning, such as image classification, it is becoming increasingly essential to tackle the vulnerability of neural networks when facing adversarial attacks, which have been widely studied recently. In the cooperative multi-agent reinforcement learning field, which has also shown potential in real-life domains, little work focuses on the problem of adversarial a ttacks. However, adversarial attacks on observations that can undermine the coordination among agents are likely to occur in actual deployment. This paper proposes a training framework that progressively generates adversarial attacks on agents' observations to help agents learn a robust cooperative policy. One attacker makes decisions on a hybrid action space that it first chooses an agent to attack and then outputs the perturbation vector. The victim policy is then trained against the attackers. Experimental results show that our generated adversarial attacks are diverse enough to improve the agents' robustness against possible disturbances.

****

Limitations of Piecewise Linearity for Efficient Robustness Certification
Klas Leino

Certified defenses against small-norm adversarial examples have received growing attention in recent years; though certified accuracies of state-of-the-art methods remain far below their non-robust counterparts, despite the fact that benchmark datasets have been shown to be well-separated at far larger radii than the literature generally attempts to certify. In this work, we offer insights that identify potential factors in this performance gap. Specifically, our analysis reveals that piecewise linearity imposes fundamental limitations on the tightness of leading certification techniques. These limitations are felt in practical terms as a greater need for capacity in models hoped to be certified efficiently. Moreover, this is _in addition_ to the capacity necessary to learn a robust boundary, studied in prior work. However, we argue that addressing the limitations of piecewise linearity through scaling up model capacity may give rise to potential difficulties---particularly regarding robust generalization---therefore, we conclude by suggesting that developing _smooth_ activation functions may be the way forward for advancing the performance of certified neural networks.

****

Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations
Xiang Fu,Zhenghao Wu,Wujie Wang,Tian Xie,Sinan Keten,Rafael Gomez-Bombarelli,Tommi S. Jaakkola

Molecular dynamics (MD) simulation techniques are widely used for various natural science applications. Increasingly, machine learning (ML) force field (FF) models begin to replace ab-initio simulations by predicting forces directly from atomic structures. Despite significant progress in this area, such techniques are primarily benchmarked by their force/energy prediction errors, even though the practical use case would be to produce realistic MD trajectories. We aim to fill this gap by introducing a novel benchmark suite for ML MD simulation. We curate representative MD systems, including water, organic molecules, peptide, and materials, and design evaluation metrics corresponding to the scientific objectives of respective systems. We benchmark a collection of state-of-the-art (SOTA) ML FF models and illustrate, in particular, how the commonly benchmarked force accuracy is not well aligned with relevant simulation metrics. We demonstrate when and how selected SOTA methods fail, along with offering directions for further improvement. Specifically, we identify stability as a key metric for ML models to improve. Our benchmark suite comes with a comprehensive open source codebase for training and simulation with ML FFs to facilitate further work.

****

FlexRound: Learnable Rounding by Element-wise Division for Post-Training Quantization
Jung Hyun Lee,Jeonghoon Kim,Se Jung Kwon,Dongsoo Lee

Post-training Quantization (PTQ) has been gaining popularity for the deployment of deep neural networks on resource-limited devices since unlike quantization-aware training, neither a full training dataset nor end-to-end training is required at all. As PTQ schemes based on reconstructing each layer or block output turn out to be effective to enhance quantized model performance, recent works have developed algorithms to devise and learn a new weight-rounding scheme so as to better reconstruct each layer or block output. We notice that, however, such new rounding schemes are established on element-wise addition. In this work, we propo se a simple yet effective new rounding mechanism for PTQ, coined FlexRound, via element-wise division to learn not only a common quantization grid size but also a different scale for each pre-trained weight. Thanks to the reciprocal rule of derivatives induced by element-wise division, FlexRound is inherently able to exploit the importance of a pre-trained weight when updating its corresponding scale, and thus, flexibly quantize a pre-trained weight depending on its own importance. We empirically validate the efficacy of FlexRound on a wide range of models and tasks. To the best of our knowledge, our work is the first to carry out comprehensive experiments on image classification, natural language understanding, and natural language generation in the per-tensor uniform PTQ setting. Our code will be open-sourced soon.

****

Re-calibrating Feature Attributions for Model Interpretation

Peiyu Yang,NAVEED AKHTAR,Zeyi Wen,Mubarak Shah,Ajmal Saeed Mian

The ability to interpret machine learning models is critical for high-stakes applications. Due to its desirable theoretical properties, path integration is a widely used scheme for feature attribution to interpret model predictions. However, the methods implementing this scheme currently rely on absolute attribution scores to eventually provide sensible interpretations. This not only contradicts the premise that the features with larger attribution scores are more relevant to the model prediction, but also conflicts with the theoretical settings for which the desirable properties of the attributions are proven. We address this by devising a method to first compute an appropriate reference for the path integration scheme. This reference further helps in identifying valid interpolation points on a desired integration path. The reference is computed in a gradient ascending direction on the model's loss surface, while the interpolations are performed by analyzing the model gradients and variations between the reference and the input. The eventual integration is effectively performed along a non-linear path. Our scheme can be incorporated into the existing integral-based attribution methods. We also devise an effective sampling and integration procedure that enables employing our scheme with multi-reference path integration efficiently. We achieve a marked performance boost for a range of integral-based attribution methods on both local and global evaluation metrics by enhancing them with our scheme. Our extensive results also show improved sensitivity, sanity preservation and model robustness with the proposed re-calibration of the attribution techniques with our method.

****

Adversarial Diversity in Hanabi

Brandon Cui,Andrei Lupu,Samuel Sokota,Hengyuan Hu,David J Wu,Jakob Nicolaus Foerster

Many Dec-POMDPs admit a qualitatively diverse set of ''reasonable'' joint policies, where reasonableness is indicated by symmetry equivariance, non-sabotaging behaviour and the graceful degradation of performance when paired with ad-hoc partners. Some of the work in diversity literature is concerned with generating these policies. Unfortunately, existing methods fail to produce teams of agents that are simultaneously diverse, high performing, and reasonable. In this work, we propose a novel approach, adversarial diversity (ADVERSITY), which is designed for turn-based Dec-POMDPs with public actions. ADVERSITY relies on off-belief learning to encourage reasonableness and skill, and on ''repulsive'' fictitious transitions to encourage diversity. We use this approach to generate new agents with distinct but reasonable play styles for the card game Hanabi and open-source our agents to be used for future research on (ad-hoc) coordination.

****

3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation

Ho Hin Lee,Shunxing Bao,Yuankai Huo,Bennett A. Landman

The recent 3D medical ViTs (e.g., SwinUNETR) achieve the state-of-the-art performances on several 3D volumetric data benchmarks, including 3D medical image segmentation. Hierarchical transformers (e.g., Swin Transformers) reintroduced several ConvNet priors and further enhanced the practical viability of adapting volum etric segmentation in 3D medical datasets. The effectiveness of hybrid approaches is largely credited to the large receptive field for non-local self-attention and the large number of model parameters. We hypothesize that volumetric ConvNets can simulate the large receptive field behavior of these learning approaches with fewer model parameters using depth-wise convolution. In this work, we propose a lightweight volumetric ConvNet, termed 3D UX-Net, which adapts the hierarchical transformer using ConvNet modules for robust volumetric segmentation. Specifically, we revisit volumetric depth-wise convolutions with large kernel (LK) size (e.g. starting from $7\times7\times7$) to enable the larger global receptive fields, inspired by Swin Transformer. We further substitute the multi-layer perceptron (MLP) in Swin Transformer blocks with pointwise depth convolutions and enhance model performances with fewer normalization and activation layers, thus reducing the number of model parameters. 3D UX-Net competes favorably with current SOTA transformers (e.g. SwinUNETR) using three challenging public datasets on volumetric brain and abdominal imaging: 1) MICCAI Challenge 2021 FLARE, 2) MICCAI Challenge 2021 FeTA, and 3) MICCAI Challenge 2022 AMOS. 3D UX-Net consistently outperforms SwinUNETR with improvement from 0.929 to 0.938 Dice (FLARE2021) and 0.867 to 0.874 Dice (Feta2021). We further evaluate the transfer learning capability of 3D UX-Net with AMOS2022 and demonstrates another improvement of 2.27% Dice (from 0.880 to 0.900). The source code with our proposed model are available at https://github.com/MASILab/3DUX-Net.

****

Push and Pull: Competing Feature-Prototype Interactions Improve Semi-supervised Semantic Segmentation

Yuhang Ding,Yifan Sun,Yi Yang

This paper challenges semi-supervised segmentation with a rethink on the feature-prototype interaction in the classification head. Specifically, we view each weight vector in the classification head as the prototype of a semantic category. The basic practice in the softmax classifier is to pull a feature towards its positive prototype (i.e., the prototype of its class), as well as to push it away from its negative prototypes. In this paper, we focus on the interaction between the feature and its negative prototypes, which is always “pushing” to make them dissimilar. While the pushing-away interaction is necessary, this paper reveals a new mechanism that the contrary interaction of pulling close negative prototypes is also beneficial. We have two insights for this counter-intuitive interaction: 1) some pseudo negative prototypes might actually be positive so that the pulling interaction can help resisting the pseudo-label noises, and 2) some true negative prototypes might contain contextual information that is beneficial. Therefore, we integrate these two competing interactions into a Push-and-Pull Learning (PPL) method. On the one hand, PPL introduces the novel pulling-close interaction between features and negative prototypes with a feature-to-prototype attention. On the other hand, PPL reinforces the original pushing-away interaction with a multi-prototype contrastive learning. While PPL is very simple, experiments show that it substantially improves semi-supervised segmentation and sets a new state of the art.

****

Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small

Kevin Ro Wang,Alexandre Variengien,Arthur Conmy,Buck Shlegeris,Jacob Steinhardt

Research in mechanistic interpretability seeks to explain behaviors of ML models in terms of their internal components. However, most previous work either focuses on simple behaviors in small models, or describes complicated behaviors in larger models with broad strokes. In this work, we bridge this gap by presenting an explanation for how GPT-2 small performs a natural language task that requires logical reasoning: indirect object identification (IOI). Our explanation encompasses 28 attention heads grouped into 7 main classes, which we discovered using a combination of interpretability approaches including causal interventions and projections.

To our knowledge, this investigation is the largest end-to-end attempt at reverse-engineering a natural behavior "in the wild" in a language model. We evaluate

the reliability of our explanation using three quantitative criteria - faithfulness, completeness and minimality. Though these criteria support our explanation, they also point to remaining gaps in our understanding.
Our work provides evidence that a mechanistic understanding of large ML models is feasible, opening opportunities to scale our understanding to both larger models and more complex tasks.
****
Equivariant Descriptor Fields: SE(3)-Equivariant Energy-Based Models for End-to-End Visual Robotic Manipulation Learning
Hyunwoo Ryu,Hong-in Lee,Jeong-Hoon Lee,Jongeun Choi
End-to-end learning for visual robotic manipulation is known to suffer from sample inefficiency, requiring large numbers of demonstrations. The spatial roto-translation equivariance, or the SE(3)-equivariance can be exploited to improve the sample efficiency for learning robotic manipulation. In this paper, we present SE(3)-equivariant models for visual robotic manipulation from point clouds that can be trained fully end-to-end. By utilizing the representation theory of the Lie group, we construct novel SE(3)-equivariant energy-based models that allow highly sample efficient end-to-end learning. We show that our models can learn from scratch without prior knowledge and yet are highly sample efficient (5~10 demonstrations are enough). Furthermore, we show that our models can generalize to tasks with (i) previously unseen target object poses, (ii) previously unseen target object instances of the category, and (iii) previously unseen visual distractors. We experiment with 6-DoF robotic manipulation tasks to validate our models' sample efficiency and generalizability. Codes are available at: https://github.com/tomato1mule/edf
****
Anatomical Structure-Aware Image Difference Graph Learning for Difference-Aware Medical Visual Question Answering
Xinyue Hu,Lin Gu,Qingyu Chen,Liangchen Liu,Kazuma Kobayashi,Qiyuan An,Zhang Mengliang,Tatsuya Harada,Zhiyong Lu,Yingying Zhu
To contribute to automating the medical vision-language model, we propose a novel Chest-Xray Different Visual Question Answering (VQA) task. Given a pair of main and reference images, this task attempts to answer several questions on both diseases and more importantly, the differences between them. This is consistent with the radiologist's diagnosis practice that compares the current image with the reference before concluding the report. For this task, we propose a new dataset, namely MIMIC-Diff-VQA including 700,821 QA pairs on 109,872 pairs of images. Meanwhile, we also propose a novel expert knowledge-aware graph representation learning model to address this problem. We leveraged the expert knowledge such as anatomical structure prior, semantic and spatial knowledge to construct a multi-relationship graph to represent the image differences between two images for the image difference VQA task. Our dataset and code will be released upon publication. We believe this work would further push forward the medical vision language model.
****
Explaining Temporal Graph Models through an Explorer-Navigator Framework
Wenwen Xia,Mincai Lai,Caihua Shan,Yao Zhang,Xinnan Dai,Xiang Li,Dongsheng Li
While GNN explanation has recently received significant attention, existing works are consistently designed for static graphs. Due to the prevalence of temporal graphs, many temporal graph models have been proposed, but explaining their predictions remains to be explored. To bridge the gap, in this paper, we propose T-GNNExplainer for temporal graph model explanation. Specifically, we regard a temporal graph constituted by a sequence of temporal events. Given a target event, our task is to find a subset of previously occurred events that lead to the model's prediction for it. To handle this combinatorial optimization problem, T-GNNExplainer includes an explorer to find the event subsets with Monte Carlo Tree Search (MCTS) and a navigator that learns the correlations between events and helps reduce the search space. In particular, the navigator is trained in advance and then integrated with the explorer to speed up searching and achieve better results. To the best of our knowledge, T-GNNExplainer is the first explainer tailo

red for temporal graph models. We conduct extensive experiments to evaluate the performance of T-GNNExplainer. Experimental results on both real-world and synthetic datasets demonstrate that T-GNNExplainer can achieve superior performance with up to about 50% improvement in Area under Fidelity-Sparsity Curve.

****

Tackling Diverse Tasks via Cross-Modal Transfer Learning
Junhong Shen,Liam Li,Lucio M. Dery,Corey Staten,Mikhail Khodak,Graham Neubig,Ameet Talwalkar

Fine-tuning large-scale pretrained models has led to remarkable progress in well-studied modalities such as vision and NLP. However, similar gains have not been observed in many other tasks due to an assumed lack of relevant pretrained models for these diverse modalities. In this work, we revisit this assumption by studying the cross-modal transfer ability of large-scale pretrained models. We introduce ORCA, a general cross-modal fine-tuning workflow that enables fast and automatic exploitation of existing pretrained models for diverse tasks. ORCA achieves task-specific adaptation by learning feature embeddings that minimize an optimal transport distance metric to map the data distribution in the end-task modality to the pretraining modality. We test ORCA on 13 tasks with varying modalities and input-output types. ORCA performs the best on 10 of them and is in the top three on the others. We further quantify the importance of embedding distance for downstream performance, highlight ORCA’s utility for data-limited tasks, and demonstrate its compatibility with same-modality transfer.

****

Leveraged Asymmetric Loss with Disambiguation for Multi-label Recognition with One-Positive Annotations
Jingyi Cui,Tao Huang,Hanyuan Hang,Yisen Wang,James Kwok

In the problem of multi-label learning from single positive labels (SPL), we learn the potential multiple labels from one observable single positive annotation. Despite many efforts to solve this problem, an effective algorithm with sound theoretical understanding is still in need. In this paper, we propose a novel loss function for the SPL problem, called leveraged asymmetric loss with disambiguation (LASD), where we introduce a pair of leverage parameters to address the severe negative-positive imbalance. From the theoretical perspective, we analyze the SPL problem, for the first time, from the perspective of risk consistency, which links the SPL loss with losses for ordinary multi-label classification. We prove the consistency of our proposed LASD loss to the cost-sensitive Hamming loss, which provides guidance to the empirical choice of our proposed leverage parameters. In experiments, we demonstrate the effectiveness of our proposed LASD loss function over other state-of-the-art methods and empirically verify our theoretical results.

****

Safe Reinforcement Learning with Contrastive Risk Prediction
Hanping Zhang,Yuhong Guo

As safety violations can lead to severe consequences in real-world applications, the increasing deployment of Reinforcement Learning (RL) in safety-critical domains such as robotics has propelled the study of safe exploration for reinforcement learning (safe RL). In this work, we propose a risk preventive training method for safe RL, which learns a statistical contrastive classifier to predict the probability of a state-action pair leading to unsafe states. Based on the predicted risk probabilities, we can collect risk preventive trajectories and reshape the reward function with risk penalties to induce safe RL policies. We conduct experiments in robotic simulation environments. The results show the proposed approach has comparable performance with the state-of-the-art model-based methods and outperforms conventional model-free safe RL approaches.

****

Analysis of differentially private synthetic data: a general measurement error approach
Yangdi Jiang,Yi Liu,Xiaodong Yan,Anne-Sophie Charest,Linglong Kong,Bei Jiang

Differential private (DP) synthetic datasets have been receiving significant att ention from academia, industry, and government. However, little is known about how to perform statistical inference using DP synthetic datasets. Naive approaches that do not take into account the induced uncertainty due to DP mechanism will result in biased estimators and invalid inferences. In this paper, we present a general class of bias-corrected DP estimators with valid asymptotic confidence intervals for parameters in regression settings, by establishing the connection between additive DP mechanisms and measurement error models. Our simulation shows that when the sample covariance between DP noises and data is close to zero, our estimator is far superior to the widely used sufficient statistic perturbation algorithm, and the CIs can achieve better coverage when comparing to the naive CIs obtained from ignoring the DP mechanism.

****

On the Efficacy of Server-Aided Federated Learning against Partial Client Participation

Haibo Yang,Peiwen Qiu,Prashant Khanduri,Jia Liu

Although federated learning (FL) has become a prevailing distributed learning framework in recent years due to its benefits in scalability/privacy, there remain many significant challenges in FL system design. Notably, most existing works in the current FL literature assume either full client or uniformly distributed client participation. Unfortunately, this idealistic assumption rarely hold in practice. It has been frequently observed that some clients may never participate in FL training (aka partial/incomplete participation) due to a meld of system heterogeneity factors. To mitigate impacts of partial client participation, an increasingly popular approach in practical FL systems is the sever-aided federated learning (SA-FL) framework, where one equips the server with an auxiliary dataset. However, despite the fact that SA-FL has been empirically shown to be effective in addressing the partial client participation problem, there remains a lack of theoretical understanding for SA-FL. Worse yet, even the ramifications of partial worker participation is not clearly understood in conventional FL so far. These theoretical gaps motivate us to rigorously investigate SA-FL. To this end, we first reveal that conventional FL is {\em not} PAC-learnable under partial participation in the worst case, which advances our understanding of conventional FL. Then, we show that the PAC-learnability of FL with partial client participation can indeed be revived by SA-FL, which theoretically justifies the use of SA-FL for the first time. Lastly, to further make SA-FL communication-efficient, we propose the \alg (\ul{s}erver-\ul{a}ided \ul{f}ederated \ul{a}ve\ul{r}ag\ul{i}ng) algorithm that enjoys convergence guarantee and the same level of communication efficiency and privacy as state-of-the-art FL.

****

Soft Neighbors are Positive Supporters in Contrastive Visual Representation Learning

Chongjian GE,Jiangliu Wang,Zhan Tong,Shoufa Chen,Yibing Song,Ping Luo

Contrastive learning methods train visual encoders by comparing views (e.g., often created via a group of data augmentations on the same instance) from one instance to others. Typically, the views created from one instance are set as positive, while views from other instances are negative. This binary instance discrimination is studied extensively to improve feature representations in self-supervised learning. In this paper, we rethink the instance discrimination framework and find the binary instance labeling insufficient to measure correlations between different samples. For an intuitive example, given a random image instance, there may exist other images in a mini-batch whose content meanings are the same (i.e., belonging to the same category) or partially related (i.e., belonging to a similar category). How to treat the images that correlate similarly to the current image instance leaves an unexplored problem. We thus propose to support the current image by exploring other correlated instances (i.e., soft neighbors). We first carefully cultivate a candidate neighbor set, which will be further utilized to explore the highly-correlated instances. A cross-attention module is then introduced to predict the correlation score (denoted as positiveness) of other correlated instances with respect to the current one. The positiveness score quantitatively measures the positive support from each correlated instance, and is e ncoded into the objective for pretext training. To this end, our proposed method benefits in discriminating uncorrelated instances while absorbing correlated instances for SSL. We evaluate our soft neighbor contrastive learning method (SNCLR) on standard visual recognition benchmarks, including image classification, object detection, and instance segmentation. The state-of-the-art recognition performance shows that SNCLR is effective in improving feature representations from both ViT and CNN encoders.

****

LA-BALD: An Information-Theoretic Image Labeling Task Sampler

Yuan-Hong Liao,Sanja Fidler

Large-scale visual recognition datasets with high-quality labels enable many computer vision applications, but also come with enormous annotation costs, especially since multiple annotators are typically queried per image to obtain a more reliable label. Recent work in label aggregation consolidates human annotations by combining them with the predictions of an online-learned predictive model. In this work, we devise an image labeling task sampler that actively selects image-worker pairs to efficiently reduce the noise in the human annotations and improve the predictive model at the same time. We propose an information-theoretic task sampler, Label Aggregation BALD (LA-BALD), to maximize the information contributing to the labeled dataset via human annotations and the model. The simulated experiments on ImageNet100-sandbox show that LA-BALD reduces the number of annotations by 19% and 12% on average compared to the two types of baselines. Our analysis shows that LA-BALD provides both more accurate annotations and a better online-learned predictive model, leading to better labeling efficiency over the baselines.

****

Offline RL for Natural Language Generation with Implicit Language Q Learning

Charlie Victor Snell,Ilya Kostrikov,Yi Su,Sherry Yang,Sergey Levine

Large language models distill broad knowledge from text corpora. However, they can be inconsistent when it comes to completing user specified tasks. This issue can be addressed by finetuning such models via supervised learning on curated datasets, or via reinforcement learning. In this work, we propose a novel offline RL method, implicit language Q-learning (ILQL), designed for use on language models, that combines both the flexible utility maximization framework of RL algorithms with the ability of supervised learning to leverage previously collected data, as well as its simplicity and stability. Our method employs a combination of value conservatism alongside an implicit dataset support constraint in learning value functions, which are then used to guide language model generations towards maximizing user-specified utility functions. In addition to empirically validating ILQL, we present a detailed empirical analysis of situations where offline RL can be useful in natural language generation settings, demonstrating how it can be a more effective utility optimizer than prior approaches for end-to-end dialogue, and how it can effectively optimize high variance reward functions based on subjective judgement, such as whether to label a comment as toxic or not.

****

MoCa: Cognitive Scaffolding for Language Models in Causal and Moral Judgment Tasks

Allen Nie,Yuhui Zhang,Atharva Amdekar,Christopher J Piech,Tatsunori Hashimoto,Tobias Gerstenberg

Human commonsense understanding of the physical and social world is organized around intuitive theories. These theories support making causal and moral judgments. When something bad happened, we naturally ask: who did what, and why? A rich literature in cognitive science has studied people's causal and moral intuitions. These works have revealed a number of factors that systematically influence people's judgments, such as the presence of norms, and whether or not the protagonist in a scenario was aware of their action's potential consequences. Here, we investigate whether large language models (LLMs) make causal and moral judgments about text-based scenarios that align with those of human participants. We find that without any annotations, LLMs and human participants are not well aligned (

17\%-39\% agreement). However, LLMs can accurately annotate what relevant factors are present in a scenario with simple expert-written instructions. We demonstrate how these annotations can be used to bring LLMs in closer alignment with people (36.3\%-47.2\% agreement). These results show how insights from cognitive science can help scaffold language models to more closely match human intuitions in challenging commonsense evaluation tasks.

****

Anchor Sampling for Federated Learning with Partial Client Participation

Feijie Wu,Song Guo,Zhihao Qu,Shiqi He,Ziming Liu

In federated learning, the support of partial client participation offers a flexible training strategy, but it deteriorates the model training efficiency. In this paper, we propose a framework FedAMD to improve the convergence property and maintain flexibility. The core idea is anchor sampling, which disjoints the partial participants into anchor and miner groups. Each client in the anchor group aims at the local bullseye with the gradient computation using a large batch. Guided by the bullseyes, clients in the miner group steer multiple near-optimal local updates using small batches and update the global model. With the joint efforts from both groups, FedAMD is able to accelerate the training process as well as improve the model performance. Measured by ϵ-approximation and compared to the state-of-the-art first-order methods, FedAMD achieves the convergence by up to $O(1/\epsilon)$ fewer communication rounds under non-convex objectives. In specific, we achieve a linear convergence rate under PL conditions. Empirical studies on real-world datasets validate the effectiveness of FedAMD and demonstrate the superiority of our proposed algorithm: Not only does it considerably save computation and communication costs, but also the test accuracy significantly improves.

****

Lattice Convolutional Networks for Learning Ground States of Quantum Many-Body Systems

Cong Fu,Xuan Zhang,Huixin Zhang,Hongyi Ling,Shenglong Xu,Shuiwang Ji

Deep learning methods have been shown to be effective in representing ground-state wave functions of quantum many-body systems. Existing methods use convolutional neural networks (CNNs) for square lattices due to their image-like structures. For non-square lattices, existing method uses graph neural network (GNN) in which structure information is not precisely captured, thereby requiring additional hand-crafted sublattice encoding. In this work, we propose lattice convolutions in which a set of proposed operations are used to convert non-square lattices into grid-like augmented lattices on which regular convolution can be applied. Based on the proposed lattice convolutions, we design lattice convolutional networks (LCN) that use self-gating and attention mechanisms. Experimental results show that our method achieves performance on par or better than the GNN method on spin 1/2 J_1-J_2 Heisenberg model over the square, honeycomb, triangular, and kagome lattices while without using hand-crafted encoding.

****

CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos

Hao-Wen Dong,Naoya Takahashi,Yuki Mitsufuji,Julian McAuley,Taylor Berg-Kirkpatrick

Recent years have seen progress beyond domain-specific sound separation for speech or music towards universal sound separation for arbitrary sounds. Prior work on universal sound separation has investigated separating a target sound out of an audio mixture given a text query. Such text-queried sound separation systems provide a natural and scalable interface for specifying arbitrary target sounds. However, supervised text-queried sound separation systems require costly labeled audio-text pairs for training. Moreover, the audio provided in existing datasets is often recorded in a controlled environment, causing a considerable generalization gap to noisy audio in the wild. In this work, we aim to approach text-queried universal sound separation by using only unlabeled data. We propose to leverage the visual modality as a bridge to learn the desired audio-textual correspondence. The proposed CLIPSep model first encodes the input query into a query vector using the contrastive language-image pretraining (CLIP) model, and the que

ry vector is then used to condition an audio separation model to separate out the target sound. While the model is trained on image-audio pairs extracted from unlabeled videos, at test time we can instead query the model with text inputs in a zero-shot setting, thanks to the joint language-image embedding learned by the CLIP model. Further, videos in the wild often contain off-screen sounds and background noise that may hinder the model from learning the desired audio-textual correspondence. To address this problem, we further propose an approach called noise invariant training for training a query-based sound separation model on noisy data. Experimental results show that the proposed models successfully learn text-queried universal sound separation using only noisy unlabeled videos, even achieving competitive performance against a supervised model in some settings.
****
On the Soft-Subnetwork for Few-Shot Class Incremental Learning
Haeyong Kang,Jaehong Yoon,Sultan Rizky Hikmawan Madjid,Sung Ju Hwang,Chang D. Yoo
Inspired by Regularized Lottery Ticket Hypothesis, which states that competitive smooth (non-binary) subnetworks exist within a dense network, we propose a few-shot class-incremental learning method referred to as Soft-SubNetworks (SoftNet). Our objective is to learn a sequence of sessions incrementally, where each session only includes a few training instances per class while preserving the knowledge of the previously learned ones. SoftNet jointly learns the model weights and adaptive non-binary soft masks at a base training session in which each mask consists of the major and minor subnetwork; the former aims to minimize catastrophic forgetting during training, and the latter aims to avoid overfitting to a few samples in each new training session. We provide comprehensive empirical validations demonstrating that our SoftNet effectively tackles the few-shot incremental learning problem by surpassing the performance of state-of-the-art baselines over benchmark datasets.
****
Approximating How Single Head Attention Learns
Charlie Victor Snell,Ruiqi Zhong,Dan Klein,Jacob Steinhardt
Why do models often attend to salient words, and how does this evolve throughout training? We approximate model training as a two stage process: early on in training when the attention weights are uniform, the model learns to translate individual input word `i` to `o` if they co-occur frequently. Later, the model learns to attend to `i` while the correct output is o because it knows `i` translates to `o`. To formalize, we define a model property, Knowledge to Translate Individual Words (KTIW) (e.g. knowing that `i` translates to `o`), and claim that it drives the learning of the attention. This claim is supported by the fact that before the attention mechanism is learned, KTIW can be learned from word co-occurrence statistics, but not the other way around. Particularly, we can construct a training distribution that makes KTIW hard to learn, the learning of the attention fails, and the model cannot even learn the simple task of copying the input words to the output. Our approximation explains why models sometimes attend to salient words, and inspires a toy example where a multi-head attention model can overcome the above hard training distribution by improving learning dynamics rather than expressiveness. We end by discussing the limitation of our approximation framework and suggest future directions.
****
Efficient Attention via Control Variates
Lin Zheng,Jianbo Yuan,Chong Wang,Lingpeng Kong
Random-feature-based attention (RFA) is an efficient approximation of softmax attention with linear runtime and space complexity. However, the approximation gap between RFA and conventional softmax attention is not well studied. Built upon previous progress of RFA, we characterize this gap through the lens of control variates and show that RFA can be decomposed into a sum of multiple control variate estimators for each element in the sequence. This new framework reveals that exact softmax attention can be recovered from RFA by manipulating each control variate. Besides, it allows us to develop a more flexible form of control variates, resulting in a novel attention mechanism that significantly reduces the appro

ximation gap while maintaining linear complexity. Extensive experiments demonstrate that our model outperforms state-of-the-art efficient attention mechanisms on both vision and language tasks.

****

Learning to Optimize Quasi-Newton Methods

Isaac Liao,Rumen Dangovski,Jakob Nicolaus Foerster,Marin Soljacic

We introduce a novel machine learning optimizer called LODO, which online meta-learns an implicit inverse Hessian of the loss as a subroutine of quasi-Newton optimization. Our optimizer merges Learning to Optimize (L2O) techniques with quasi-Newton methods to learn neural representations of symmetric matrix vector products, which are more flexible than those in other quasi-Newton methods. Unlike other L2O methods, ours does not require any meta-training on a training task distribution, and instead learns to optimize on the fly while optimizing on the test task, adapting to the local characteristics of the loss landscape while traversing it. Theoretically, we show that our optimizer approximates the inverse Hessian in noisy loss landscapes and is capable of representing a wide range of inverse Hessians. We experimentally verify our algorithm's performance in the presence of noise, and show that simpler alternatives for representing the inverse Hessians worsen performance. Lastly, we use our optimizer to train a semi-realistic deep neural network with 95k parameters, and obtain competitive results against standard neural network optimizers.

****

Toxicity in Multilingual Machine Translation at Scale

Marta R. Costa-jussà,Christophe Ropers,Eric Michael Smith,Daniel Licht,Carlos Escolano,Javier Ferrando

Machine Translation systems can produce different types of errors, some of which get characterized as critical or catastrophic due to the specific negative impact they can have on users. Automatic or human evaluation metrics do not necessarily differentiate between such critical errors and more innocuous ones. In this paper we focus on one type of critical error: added toxicity. We evaluate and analyze added toxicity when translating a large evaluation dataset (HOLISTICBIAS, over 472k sentences, covering 13 demographic axes) from English into 164 languages. The toxicity automatic evaluation shows that added toxicity across languages varies from 0% to 5%. The output languages with the most added toxicity tend to be low-resource ones, and the demographic axes with the most added toxicity include sexual orientation, gender and sex, and ability. We also perform human evaluation on a subset of 8 directions, confirming the prevalence of true added toxicity.

We use a measurement of the amount of source contribution to the translation, where a low source contribution implies hallucination, to interpret what causes toxicity. We observe that the source contribution is somewhat correlated with toxicity but that 45.6% of added toxic words have a high source contribution, suggesting that much of the added toxicity may be due to mistranslations. Combining the signal of source contribution level with a measurement of translation robustness allows us to flag 22.3% of added toxicity, suggesting that added toxicity may be related to both hallucination and the stability of translations in different contexts. Given these findings, our recommendations to reduce added toxicity are to curate training data to avoid mistranslations, mitigate hallucination and check unstable translations.

****

An Adaptive Policy to Employ Sharpness-Aware Minimization

Weisen Jiang,Hansi Yang,Yu Zhang,James Kwok

Sharpness-aware minimization (SAM), which searches for flat minima by min-max optimization, has been shown to be useful in improving model generalization. However, since each SAM update requires computing two gradients, its computational cost and training time are both doubled compared to standard empirical risk minimization (ERM). Recent state-of-the-arts reduce the fraction of SAM updates and thus accelerate SAM by switching between SAM and ERM updates randomly or periodically. In this paper, we design an adaptive policy to employ SAM based on the loss landscape geometry. Two efficient algorithms, AE-SAM and AE-LookSAM, are proposed. We theoretically show that AE-SAM has the same convergence rate as SAM. Experimental results on various datasets and architectures demonstrate the efficiency and effectiveness of the adaptive policy.

****

FedMT: Federated Learning with Mixed-type Labels

Qiong Zhang,Aline Talhouk,Gang Niu,Xiaoxiao Li

In federated learning (FL), classifiers (e.g., deep networks) are trained on datasets from multiple centers without exchanging data across them, and thus improves sample efficiency. In the classical setting of FL, the same labeling criterion is usually employed across all centers being involved in training. This constraint greatly limits the applicability of FL. For example, standards used for disease diagnosis are more likely to be different across clinical centers, which mismatches the classical FL setting. In this paper, we consider an important yet under-explored setting of FL, namely FL with mixed-type labels where different labeling criteria can be employed by various centers, leading to inter-center label space differences and challenging existing FL methods designed for the classical setting. To effectively and efficiently train models with mixed-type labels, we propose a theory-guided and model-agnostic approach that can make use of the underlying correspondence between those label spaces and can be easily combined with various FL methods such as FedAvg. We present convergence analysis based on over-parameterized ReLU networks. We show that the proposed method can achieve linear convergence in label projection, and demonstrate the impact of the parameters of our new setting on the convergence rate. The proposed method is evaluated and the theoretical findings are validated on benchmark and medical datasets.

****

Penalizing the High-likelihood: A Novel Sampling Method for Open-ended Neural Text Generation via Inverse Probability Weighting

Xinran Zhang,Maosong Sun,Jiafeng Liu,Xiaobing Li

Traditional stochastic sampling methods for open-ended neural text generation focus on truncating the low-likelihood part of the predicted distribution. They do not directly manipulate the high-likelihood part, which leads to the likelihood trap that induces repetition and boredom. They also do not directly leverage that human does not always favor high-likelihood texts. Inspired by these, we propose a novel sampling method that rescales the high-likelihood part of the distribution with inverse probability weighting. It increases the diversity by rescaling and penalizing the high-likelihood words, and preserves the fluency by using multi-filtering truncation on the low-likelihood words. We use pre-trained language models to compare our algorithm with traditional sampling methods. Results show that our algorithm can significantly increase the diversity and novelty of generated texts without corrupting the fluency.

****

Optimal Conservative Offline RL with General Function Approximation via Augmented Lagrangian

Paria Rashidinejad,Hanlin Zhu,Kunhe Yang,Stuart Russell,Jiantao Jiao

Offline reinforcement learning (RL), which aims at learning good policies from historical data, has received significant attention over the past years. Much effort has focused on improving offline RL practicality by addressing the prevalent issue of partial data coverage through various forms of conservative policy learning. While the majority of algorithms do not have finite-sample guarantees, several provable conservative offline RL algorithms are designed and analyzed within the single-policy concentrability framework that handles partial coverage. Yet, in the nonlinear function approximation setting where confidence intervals are difficult to obtain, existing provable algorithms suffer from computational intractability, prohibitively strong assumptions, and suboptimal statistical rates. In this paper, we leverage the marginalized importance sampling (MIS) formulation of RL and present the first set of offline RL algorithms that are statistically optimal and practical under general function approximation and single-policy concentrability, bypassing the need for uncertainty quantification. We identify that the key to successfully solving the sample-based appr oximation of the MIS problem is ensuring that certain occupancy validity constraints are nearly satisfied. We enforce these constraints by a novel application of the augmented Lagrangian method and prove the following result: with the MIS formulation, augmented Lagrangian is enough for statistically optimal offline RL. In stark contrast to prior algorithms that induce additional conservatism through methods such as behavior regularization, our approach provably eliminates this need and reinterprets regularizers as "enforcers of occupancy validity" than "promoters of conservatism."

****

Bandit Learning with General Function Classes: Heteroscedastic Noise and Variance-dependent Regret Bounds

Heyang Zhao,Dongruo Zhou,Jiafan He,Quanquan Gu

We consider learning a stochastic bandit model, where the reward function belongs to a general class of uniformly bounded functions, and the additive noise can be heteroscedastic. Our model captures contextual linear bandits and generalized linear bandits as special cases. While previous works (Kirschner and Krause, 2018; Zhou et al., 2021) based on weighted ridge regression can deal with linear bandits with heteroscedastic noise, they are not directly applicable to our general model due to the curse of nonlinearity. In order to tackle this problem, we propose a \emph{multi-level learning} framework for the general bandit model. The core idea of our framework is to partition the observed data into different levels according to the variance of their respective reward and perform online learning at each level collaboratively. Under our framework, we first design an algorithm that constructs the variance-aware confidence set based on empirical risk minimization and prove a variance-dependent regret bound. For generalized linear bandits, we further propose an algorithm based on follow-the-regularized-leader (FTRL) subroutine and online-to-confidence-set conversion, which can achieve a tighter variance-dependent regret under certain conditions.

****

Injecting Image Details into CLIP's Feature Space

Zilun Zhang,Cuifeng Shen,Shen Yuan,Huixin Xiong,Xinyu Zhou

Although CLIP-like Visual Language Models provide a functional joint feature space for image and text, due to the limitation of the CILP-like model's image input size (e.g., 224), subtle details are lost in the feature representation if we input high-resolution images (e.g., 2240). In this work, we introduce an efficient framework that can produce a single feature representation for a high-resolution image that injects image details and shares the same semantic space as the original CLIP. In the framework, we train a feature fusing model based on CLIP features extracted from a carefully designed image patch method (Complete Cover) that can cover objects of any scale, weakly supervised by image-agnostic class prompted queries. We validate our framework by retrieving images from class prompted queries on the existing real-world and synthetic datasets, showing significant performance improvement on these tasks. Furthermore, to fully demonstrate our framework's detail retrieval ability, we construct a CLEVR-like synthetic dataset called CLVER-DS, which is fully annotated and has a controllable object scale.

****

Adaptive Sparse Softmax: An Effective and Efficient Softmax Variant for Text Classification

Qi Lv,Lei Geng,Ziqiang Cao,Min Cao,Sujian Li,Wenjie Li,Guohong Fu

Softmax with the cross entropy loss is the standard configuration for current neural text classification models. The gold score for a target class is supposed to be 1, but it is never reachable under the softmax schema. Such a problem makes the training process continue forever and leads to overfitting. Moreover, the “target-approach-1” training goal forces the model to continuously learn all samples, leading to a waste of time in handling some samples which have already been classified correctly with high confidence, while the test goal simply requires the target class of each sample to hold the maximum score. To solve the above weaknesses, we propose the \textbf{A}daptive \textbf{S}parse softmax (AS-Softmax) which designs a reasonable and test-matching transformation on top of softmax. F or more purposeful learning, we discard the classes with far smaller scores compared with the actual class during training. Then the model could focus on learning to distinguish the target class from its strong opponents, which is also the great challenge in test. In addition, since the training losses of easy samples will gradually drop to 0 in AS-Softmax, we develop an adaptive gradient accumulation strategy based on the masked sample ratio to speed up training. We verify proposed AS-Softmax on a variety of multi-class, multi-label and token classification tasks with class sizes ranging from 5 to 5000+. The results show that AS-Softmax consistently outperforms softmax and its variants, and the loss of AS-Softmax is remarkably correlated with classification performance in validation. Furthermore, adaptive gradient accumulation strategy can bring about 1.2× training speedup comparing with the standard softmax while maintaining classification effectiveness.

****

Stochastic Bridges as Effective Regularizers for Parameter-Efficient Tuning

Weize Chen,Xu Han,Yankai Lin,Zhiyuan Liu,Maosong Sun,Jie Zhou

Parameter-efficient tuning methods (PETs) have achieved promising results in tuning large pre-trained language models (PLMs). By formalizing frozen PLMs and additional tunable parameters as systems and controls respectively, PETs can be theoretically grounded to optimal control and further viewed as optimizing terminal cost and running cost in the optimal control literature. Despite the elegance of this theoretical grounding, in practice, existing PETs often ignore the running cost and only optimize the terminal cost, i.e., focus on optimizing the loss function of the output state, regardless of the running cost that depends on the intermediate states. Since it is non-trivial to directly model the intermediate states and design a running cost function, we propose to use latent stochastic bridges to regularize the intermediate states and serve as the running cost of PETs. As the first work to propose regularized PETs that use stochastic bridges as the regularizers (running costs) for intermediate states, we show the effectiveness and generality of this regularization across different tasks, PLMs and PETs. In view of the great potential and capacity, we believe more sophisticated regularizers can be designed for PETs and better performance can be achieved in the future.

****

Continuous Goal Sampling: A Simple Technique to Accelerate Automatic Curriculum Learning

Mehul Damani,Lerrel Pinto

Goal-conditioned reinforcement learning (RL) tackles the problem of training an RL agent to reach multiple goals in an environment, often with sparse rewards only administered upon reaching the goal.

In this regard, automatic curriculum learning can improve an agent's learning by sampling goals in a structured order catered to the agent's current ability.

This work presents two contributions to improve learning in goal-conditioned RL environments.

First, we present a simple, algorithm-agnostic technique to accelerate learning by continuous goal sampling, in which an agent's goals are sampled and changed multiple times within a single episode.

Such continuous goal sampling enables faster exploration of the goal space and allows curriculum methods to have a more significant impact on an agent's learning.

Second, we propose VDIFF, an automatic curriculum learning method that uses an agent's value function to create a self-paced curriculum by sampling goals on which the agent is demonstrating high learning progress.

Through results on 17 multi-goal robotic environments and navigation tasks, we show that continuous goal sampling and VDIFF work synergistically and result in performance gains over current state-of-the-art methods.

****

What do Vision Transformers Learn? A Visual Exploration

Amin Ghiasi,Hamid Kazemi,Steven Reich,Eitan Borgnia,Manli Shu,Micah Goldblum,Andrew Gordon Wilson,Tom Goldstein

Vision transformers (ViTs) are quickly becoming the de-facto architecture for computer vision, yet we understand very little about why they work and what they learn. While existing studies visually analyze the mechanisms of convolutional neural networks, an analogous exploration of ViTs remains challenging. In this paper, we first address the obstacles to performing visualizations on ViTs. Assisted by these solutions, we observe that neurons in ViTs trained with language model supervision (e.g., CLIP) are activated by semantic concepts rather than visual features. We also explore the underlying differences between ViTs and CNNs, and we find that transformers detect image background features, just like their convolutional counterparts, but their predictions depend far less on high-frequency information. On the other hand, both architecture types behave similarly in the way features progress from abstract patterns in early layers to concrete objects in late layers. In addition, we show that ViTs maintain spatial information in all layers except the final layer. In contrast to previous works, we show that the last layer most likely discards the spatial information and behaves as a learned global pooling operation. Finally, we conduct large-scale visualizations on a wide range of ViT variants, including DeiT, CoaT, ConViT, PiT, Swin, and Twin, to validate the effectiveness of our method.

****

Detecting and Mitigating Indirect Stereotypes in Word Embeddings

Erin George,Deanna Needell

Societal biases in the usage of words, including harmful stereotypes, are frequently learned by common word embedding methods. These biases manifest not only between a word and an explicit marker of its stereotype, but also between words that share related stereotypes. This latter phenomenon, sometimes called ``indirect bias,'' has resisted prior attempts at debiasing. In this paper, we propose a novel method to mitigate indirect bias in distributional word embeddings by modifying biased relationships between words before embeddings are learned. This is done by considering how the co-occurrence probability of a given pair of words changes in the presence of words marking an attribute of bias, and using this to average out the effect of a bias attribute. To evaluate this method, we perform a series of common tests and demonstrate that the semantic quality of the word embeddings is retained while measures of bias in the embeddings are reduced. In addition, we conduct novel tests for measuring indirect stereotypes by extending the Word Embedding Association Test (WEAT) with new test sets for indirect binary gender stereotypes. With these tests, we demonstrate that this method can reduce the presence of more subtle stereotypes not properly addressed by previous work.

****

OCIM : Object-centric Compositional Imagination for Visual Abstract Reasoning

Rim Assouel,Pau Rodriguez,Perouz Taslakian,David Vazquez,Yoshua Bengio

A long-sought property of machine learning systems is the ability to compose learned concepts in novel ways that would enable them to make sense of new situations. Such capacity for imagination -- a core aspect of human intelligence -- is not yet attained for machines. In this work, we show that object-centric inductive biases can be leveraged to derive an imagination-based learning framework that achieves compositional generalization on a series of tasks. Our method, denoted Object-centric Compositional IMagination (OCIM), decomposes visual reasoning tasks into a series of primitives applied to objects without using a domain-specific language. We show that these primitives can be recomposed to generate new imaginary tasks. By training on such imagined tasks, the model learns to reuse the previously-learned concepts to systematically generalize at test time. We test our model on a series of arithmetic tasks where the model has to infer the sequence of operations (programs) applied to a series of inputs. We find that imagination is key for the model to find the correct solution for unseen combinations of operations.

****

How Weakly Supervised Information helps Contrastive Learning

Jingyi Cui,Weiran Huang,Yisen Wang

Contrastive learning has shown outstanding performances in both supervised and u nsupervised learning. However, little is known about when and how weakly supervised information helps improve contrastive learning, especially from the theoretical perspective. The major challenge is that the existing theory of contrastive learning based on supervised learning frameworks failed to distinguish between supervised and unsupervised contrastive learning. Therefore, we turn to the unsupervised learning frameworks, and based on the posterior probability of labels, we translate the weakly supervised information into a similarity graph under the framework of spectral clustering. In this paper, we investigate two typical weakly supervised learning problems, noisy label learning, and semi-supervised learning, and analyze their influence on contrastive learning within a unified framework. Specifically, we analyze the effect of weakly supervised information on the augmentation graph of unsupervised contrastive learning, and consequently on its corresponding error bound. Numerical experiments are carried out to verify the theoretical findings.

****

A computational framework to unify representation similarity and function in biological and artificial neural networks

Xuming Ran,Jie Zhang,Ziyuan Ye,Haiyan Wu,Qi Xu,Huihui Zhou,Quanying Liu

Artificial neural network (ANN) is a versatile tool to study the neural representation in the ventral visual stream, and the knowledge in neuroscience in return inspires ANN models to improve performance in the task. However, it is still unclear how to merge these two directions into a unified framework. In this study, we propose an integrated framework called Deep Autoencoder with Neural Response (DAE-NR), which incorporates information from ANN and the visual cortex to achieve better image reconstruction performance and higher neural representation similarity between biological and artificial neurons. The same visual stimuli (i.e., natural images) are input to both the mice brain and DAE-NR. The encoder of DAE-NR jointly learns the dependencies from neural spike encoding and image reconstruction. For the neural spike encoding task, the features derived from a specific hidden layer of the encoder are transformed by a mapping function to predict the ground-truth neural response under the constraint of image reconstruction. Simultaneously, for the image reconstruction task, the latent representation obtained by the encoder is assigned to a decoder to restore the original image under the guidance of neural information. In DAE-NR, the learning process of encoder, mapping function and decoder are all implicitly constrained by these two tasks. Our experiments demonstrate that if and only if with the joint learning, DAE-NRs can improve the performance of visual image reconstruction and increase the representation similarity between biological neurons and artificial neurons. The DAE-NR offers a new perspective on the integration of computer vision and neuroscience.

****

Fairness and Accuracy under Domain Generalization

Thai-Hoang Pham,Xueru Zhang,Ping Zhang

As machine learning (ML) algorithms are increasingly used in high-stakes applications, concerns have arisen that they may be biased against certain social groups. Although many approaches have been proposed to make ML models fair, they typically rely on the assumption that data distributions in training and deployment are identical. Unfortunately, this is commonly violated in practice and a model that is fair during training may lead to an unexpected outcome during its deployment. Although the problem of designing robust ML models under dataset shifts has been widely studied, most existing works focus only on the transfer of accuracy. In this paper, we study the transfer of both fairness and accuracy under domain generalization where the data at test time may be sampled from never-before-seen domains. We first develop theoretical bounds on the unfairness and expected loss at deployment, and then derive sufficient conditions under which fairness and accuracy can be perfectly transferred via invariant representation learning. Guided by this, we design a learning algorithm such that fair ML models learned with training data still have high fairness and accuracy when deployment environments change. Experiments on real-world data validate the proposed algorithm.

****

DROP: Conservative Model-based Optimization for Offline Reinforcement Learning
Jinxin Liu,Hongyin Zhang,Zifeng Zhuang,Yachen Kang,Donglin Wang,Bin Wang,Jianye HAO

In this work, we decouple the iterative (bi-level) offline RL optimization from the offline training phase, forming a non-iterative bi-level learning paradigm that avoids the iterative error propagation over two levels. Specifically, this non-iterative paradigm allows us to conduct inner-level optimization in training (ie, employing policy/value regularization), while performing outer-level optimization in testing (ie, conducting policy inference). Naturally, such paradigm raises three core questions (that are not fully answered by prior non-iterative methods): (Q1) What information should we transfer from inner-level to outer-level? (Q2) What should we pay attention to when using the transferred information in outer-level optimization? (Q3) What are the benefits of concurrently conducting outer-level optimization during testing? Motivated by model-based optimization, we proposed DROP, which fully answered the above three questions. Particularly, in inner-level, DROP decomposes offline data into multiple subsets, and learns a score model (Q1). To keep safe exploitation to score model in outer-level, we explicitly learn a behavior embedding and introduce a conservative regularization (Q2). During testing, we show that DROP permits deployment adaptation, enabling an adaptive inference across states (Q3). Empirically, we evaluate DROP on various benchmarks, showing that DROP gains comparable or better performance compared to prior offline RL methods.

****

Language Models Can Teach Themselves to Program Better
Patrick Haluptzok,Matthew Bowers,Adam Tauman Kalai

Recent Language Models (LMs) achieve breakthrough performance in code generation when trained on human-authored problems, even solving some competitive-programming problems. Self-play has proven useful in games such as Go, and thus it is natural to ask whether LMs can generate their own instructive programming problems to improve their performance. We show that it is possible for an LM to synthesize programming problems and solutions, which are filtered for correctness by a Python interpreter. The LM’s performance is then seen to improve when it is fine-tuned on its own synthetic problems and verified solutions; thus the model “improves itself” using the Python interpreter. Problems are specified formally as programming puzzles [Schuster et al. , 2021], a code-based problem format where solutions can easily be verified for correctness by execution. In experiments on publicly-available LMs, test accuracy more than doubles. This work demonstrates the potential for code LMs, with an interpreter, to generate instructive problems and improve their own performance.

****

MVP: Multi-task Supervised Pre-training for Natural Language Generation
Tianyi Tang,Junyi Li,Xin Zhao,Ji-Rong Wen

Pre-trained language models (PLMs) have achieved remarkable success in natural language generation (NLG) tasks. Up to now, most NLG-oriented PLMs are pre-trained in an unsupervised manner using the large-scale general corpus. In the meanwhile, an increasing number of models pre-trained with labeled data (i.e., “supervised pre-training”) showcase superior performance compared to unsupervised pre-trained models. Motivated by the success of supervised pre-training, we propose Multi-task superVised Pre-training (MVP) for natural language generation. We collect a large-scale natural language generation corpus, MVPCorpus, from 77 datasets over 11 diverse NLG tasks. Then we unify these examples into a general text-to-text format to pre-train the text generation model MVP in a supervised manner. For each task, we further pre-train specific soft prompts to stimulate the model’s capacity to perform a specific task. Extensive experiments have demonstrated the effectiveness and generalizability of our MVP model in a number of NLG tasks, which achieves state-of-the-art performance on 13 out of 17 datasets.

****

Learning Unified Representations for Multi-Resolution Face Recognition
Hulingxiao He,Wu Yuan,Yidian Huang,Shilong Zhao,Wen Yuan,HanQing Li

In this work, we propose Branch-to-Trunk network (BTNet), a novel representation learning method for multi-resolution face recognition. It consists of a trunk network (TNet), namely a unified encoder, and multiple branch networks (BNets), namely resolution adapters. As per the input, a resolution-specific BNet is used and the output are implanted as feature maps in the feature pyramid of TNet, at a layer with the same resolution. The discriminability of tiny faces is significantly improved, as the interpolation error introduced by rescaling, especially up-sampling, is mitigated on the inputs. With branch distillation and backward-compatible training, BTNet transfers discriminative high-resolution information to multiple branches while guaranteeing representation compatibility. Our experiments demonstrate strong performance on face recognition benchmarks, both for multi-resolution face verification and face identification, with much less computation amount and parameter storage. We establish new state-of-the-art on the challenging QMUL-SurvFace 1: N face identification task.

****

Latent Bottlenecked Attentive Neural Processes

Leo Feng,Hossein Hajimirsadeghi,Yoshua Bengio,Mohamed Osama Ahmed

Neural Processes (NPs) are popular methods in meta-learning that can estimate predictive uncertainty on target datapoints by conditioning on a context dataset. Previous state-of-the-art method Transformer Neural Processes (TNPs) achieve strong performance but require quadratic computation with respect to the number of context datapoints, significantly limiting its scalability. Conversely, existing sub-quadratic NP variants perform significantly worse than that of TNPs. Tackling this issue, we propose Latent Bottlenecked Attentive Neural Processes (LBANPs), a new computationally efficient sub-quadratic NP variant, that has a querying computational complexity independent of the number of context datapoints. The model encodes the context dataset into a constant number of latent vectors on which self-attention is performed. When making predictions, the model retrieves higher-order information from the context dataset via multiple cross-attention mechanisms on the latent vectors. We empirically show that LBANPs achieve results competitive with the state-of-the-art on meta-regression, image completion, and contextual multi-armed bandits. We demonstrate that LBANPs can trade-off the computational cost and performance according to the number of latent vectors. Finally, we show LBANPs can scale beyond existing attention-based NP variants to larger dataset settings.

****

VoLTA: Vision-Language Transformer with Weakly-Supervised Local-Feature Alignment

Shraman Pramanick,Li Jing,Sayan Nag,Jiachen Zhu,Hardik J Shah,Yann LeCun,Rama Chellappa

Vision-language pre-training (VLP) has recently proven highly effective for various uni- and multi-modal downstream applications. However, most existing end-to-end VLP methods use high-resolution image-text-box data to perform well on fine-grained region-level tasks, such as object detection, segmentation, and referring expression comprehension. Unfortunately, such high-resolution images with accurate bounding box annotations are expensive to collect and use for supervision at scale. In this work, we propose VoLTA (Vision-Language Transformer with weakly-supervised local-feature Alignment), a new VLP paradigm that only utilizes image-caption data but achieves fine-grained region-level image understanding, eliminating the use of expensive box annotations. VoLTA adopts graph optimal transport-based weakly-supervised alignment on local image patches and text tokens to germinate an explicit, self-normalized, and interpretable low-level matching criterion. In addition, VoLTA pushes multi-modal fusion deep into the uni-modal backbones during pre-training and removes fusion-specific transformer layers, further reducing memory requirements. Extensive experiments on a wide range of vision- and vision-language downstream tasks demonstrate the effectiveness of VoLTA on fine-grained applications without compromising the coarse-grained downstream performance, often outperforming methods using significantly more caption and box annotations.

****

Represent to Control Partially Observed Systems: Representation Learning with Pr ovable Sample Efficiency

Lingxiao Wang,Qi Cai,Zhuoran Yang,Zhaoran Wang

Reinforcement learning in partially observed Markov decision processes (POMDPs) faces two challenges. (i) It often takes the full history to predict the future, which induces a sample complexity that scales exponentially with the horizon. (ii) The observation and state spaces are often continuous, which induces a sample complexity that scales exponentially with the extrinsic dimension. Addressing such challenges requires learning a minimal but sufficient representation of the observation and state histories by exploiting the structure of the POMDP.

To this end, we propose a reinforcement learning algorithm named Represent to Control (RTC), which learns the representation at two levels while optimizing the policy.~(i) For each step, RTC learns to represent the state with a low-dimensional feature, which factorizes the transition kernel. (ii) Across multiple steps, RTC learns to represent the full history with a low-dimensional embedding, which assembles the per-step feature. We integrate (i) and (ii) in a unified framework that allows a variety of estimators (including maximum likelihood estimators and generative adversarial networks). For a class of POMDPs with a low-rank structure in the transition kernel, RTC attains an $O(1/\epsilon^2)$ sample complexity that scales polynomially with the horizon and the intrinsic dimension (that is, the rank). Here ϵ is the optimality gap. To our best knowledge, RTC is the first sample-efficient algorithm that bridges representation learning and policy optimization in POMDPs with infinite observation and state spaces.

****

Towards Better Selective Classification

Leo Feng,Mohamed Osama Ahmed,Hossein Hajimirsadeghi,Amir H. Abdi

We tackle the problem of Selective Classification where the objective is to achieve the best performance on a predetermined ratio (coverage) of the dataset. Recent state-of-the-art selective methods come with architectural changes either via introducing a separate selection head or an extra abstention logit. In this paper, we challenge the aforementioned methods. The results suggest that the superior performance of state-of-the-art methods is owed to training a more generalizable classifier rather than their proposed selection mechanisms. We argue that the best performing selection mechanism should instead be rooted in the classifier itself. Our proposed selection strategy uses the classification scores and achieves better results by a significant margin, consistently, across all coverages and all datasets, without any added compute cost. Furthermore, inspired by semi-supervised learning, we propose an entropy-based regularizer that improves the performance of selective classification methods. Our proposed selection mechanism with the proposed entropy-based regularizer achieves new state-of-the-art results.

****

Offline Equilibrium Finding

Shuxin Li,Xinrun Wang,Jakub Cerny,Youzhi Zhang,Pengdeng Li,Hau Chan,Bo An

Offline reinforcement learning (Offline RL) is an emerging field that has recently begun gaining attention across various application domains due to its ability to learn behavior from earlier collected datasets. Offline RL proved very successful, paving a path to solving previously intractable real-world problems, and we aim to generalize this paradigm to a multi-agent or multiplayer-game setting. To this end, we formally introduce a problem of offline equilibrium finding (OEF) and construct multiple datasets across a wide range of games using several established methods. To solve the OEF problem, we design a model-based method that can directly apply any online equilibrium finding algorithm to the OEF setting while making minimal changes. We focus on three most prominent contemporary online equilibrium finding algorithms and adapt them to the OEF setting, creating three model-based variants: OEF-PSRO and OEF-CFR, which generalize the widely-used algorithms PSRO and Deep CFR to compute Nash equilibria (NEs), and OEF-JPSRO, which generalizes the JPSRO to calculate (Coarse) Correlated equilibria ((C)CEs). We further improve their performance by combining the behavior cloning policy with the model-based policy. Extensive experimental results demonstrate the super

iority of our approach over multiple model-based and model-free offline RL algorithms and the necessity of the model-based method for solving OEF problems. We hope that our efforts may help to accelerate research in large-scale equilibrium finding.
****

ASGNN: Graph Neural Networks with Adaptive Structure

Zepeng Zhang,Songtao Lu,Zengfeng Huang,Ziping Zhao

The graph neural network (GNN) has presented impressive achievements in numerous machine learning tasks. However, many existing GNN models are shown to be extremely vulnerable to adversarial attacks, which makes it essential to build robust GNN architectures. In this work, we propose a novel interpretable message passing scheme with adaptive structure (ASMP) to defend against adversarial attacks on graph structure. Layers in ASMP are derived based on optimization steps that minimize an objective function that simultaneously learns the node feature and the graph structure. ASMP is adaptive in the sense that the message passing process in different layers is able to be carried out over different graphs. Such a property allows more fine-grained handling of the noisy graph structure and hence improves the robustness. Integrating ASMP with neural networks can lead to a new family of GNNs with adaptive structure (ASGNN). Extensive experiments on semi-supervised node classification tasks demonstrate that the proposed ASGNN outperforms the state-of-the-art GNN architectures with respect to classification performance under various graph adversarial attacks.
****

Iteratively Learning Novel Strategies with Diversity Measured in State Distances

Wei Fu,Weihua Du,Jingwei Li,Sunli Chen,Jingzhao Zhang,Yi Wu

In complex reinforcement learning (RL) problems, policies with similar rewards may have substantially different behaviors. Yet, to not only optimize rewards but also discover as many diverse strategies as possible remains a challenging problem. A natural approach to this task is constrained population-based training (PBT), which simultaneously learns a collection of policies subject to diversity constraints. However, due to the unaffordable computation cost of PBT, we adopt an alternative approach, iterative learning (IL), which repeatedly learns a single novel policy that is sufficiently different from previous ones. We first analyze these two frameworks and prove that, for any policy pool derived by PBT, we can always use IL to obtain another policy pool of the same rewards and competitive diversity scores. In addition, we also present a novel state-based diversity measure with two tractable realizations. Such a metric can impose a stronger and much smoother diversity constraint than existing action-based metrics. Combining IL and the state-based diversity measure, we develop a powerful diversity-driven RL algorithm, State-based Intrinsic-reward Policy Optimization (SIPO), with provable convergence properties. We empirically examine our algorithm in complex multi-agent environments including StarCraft Multi-Agent Challenge and Google Research Football. SIPO is able to consistently derive strategically diverse and human-interpretable policies that cannot be discovered by existing baselines.
****

Learning Kernelized Contextual Bandits in a Distributed and Asynchronous Environment

Chuanhao Li,Huazheng Wang,Mengdi Wang,Hongning Wang

Despite the recent advances in communication-efficient distributed bandit learning, most existing solutions are restricted to parametric models, e.g., linear bandits and generalized linear bandits (GLB). In comparison, kernel bandits, which search for non-parametric functions in a reproducing kernel Hilbert space (RKHS), offer higher modeling capacity. But the only existing work in distributed kernel bandits adopts a synchronous communication protocol, which greatly limits its practical use (e.g., every synchronization step requires all clients to participate and wait for data exchange).

In this paper, in order to improve the robustness against delays and unavailability of clients that are common in practice, we propose the first asynchronous solution based on approximated kernel regression for distributed kernel bandit learning. A set of effective treatments are developed to ensure approximation quali ty and communication efficiency. Rigorous theoretical analysis about the regret and communication cost is provided; and extensive empirical evaluations demonstrate the effectiveness of our solution.

****

ATTRIBUTES RECONSTRUCTION IN HETEROGENEOUS NETWORKS VIA GRAPH AUGMENTATION

yixuan Liang,yuan wan

Heterogeneous Graph Neural Networks(HGNNs), as an effective tool for mining heterogeneous graphs, have achieved remarkable performance on node classification tasks. Yet, HGNNs are limited in their mining power as they require all nodes to have complete and reliable attributes. It is usually unrealistic since the attributes of many nodes in reality are inevitably missing or defective. Existing methods usually take imputation schemes to complete missing attributes, in which topology information is ignored, leading to suboptimal performance. And some graph augmentation techniques have improved the quality of attributes, while few of them are designed for heterogeneous graphs. In this work, we study the data augmentation on heterogeneous graphs, tackling the missing and defective attributes simultaneously, and propose a novel generic architecture—Attributes Reconstruction in Heterogeneous networks via Graph Augmentation(ARHGA), including random sampling, attribute augmentation and consistency training. In graph augmentation, to ensure attributes plausible and accurate, the attention mechanism is adopted to reconstruct attributes under the guidance of the topological relationship between nodes. Our proposed architecture can be easily combined with any GNN-based heterogeneous model, and improves the performance. Extensive experiments on three benchmark datasets demonstrate the superior performance of ARHGA over strate-of-the-art baselines on semi-supervised node classification.

****

Graph Signal Sampling for Inductive One-Bit Matrix Completion: a Closed-form Solution

Chao Chen,Haoyu Geng,Gang Zeng,Zhaobing Han,Hua Chai,Xiaokang Yang,Junchi Yan

Inductive one-bit matrix completion is motivated by modern applications such as recommender systems, where new users would appear at test stage with the ratings consisting of only ones and no zeros. We propose a unified graph signal sampling framework which enjoys the benefits of graph signal analysis and processing. The key idea is to transform each user's ratings on the items to a function (signal) on the vertices of an item-item graph, then learn structural graph properties to recover the function from its values on certain vertices --- the problem of graph signal sampling. We propose a class of regularization functionals that takes into account discrete random label noise in the graph vertex domain, then develop the GS-IMC approach which biases the reconstruction towards functions that vary little between adjacent vertices for noise reduction. Theoretical result shows that accurate reconstructions can be achieved under mild conditions. For the online setting, we develop a Bayesian extension, i.e., BGS-IMC which considers continuous random Gaussian noise in the graph Fourier domain and builds upon a prediction-correction update algorithm to obtain the unbiased and minimum-variance reconstruction. Both GS-IMC and BGS-IMC have closed-form solutions and thus are highly scalable in large data. Experiments show that our methods achieve state-of-the-art performance on public benchmarks.

****

DocPrompting: Generating Code by Retrieving the Docs

Shuyan Zhou,Uri Alon,Frank F. Xu,Zhengbao Jiang,Graham Neubig

Publicly available source-code libraries are continuously growing and changing. This makes it impossible for models of code
to keep current with all available APIs by simply training these models on existing code repositories. Thus, existing models inherently cannot generalize to using unseen functions and libraries, because these would never appear in the training data. In contrast, when human programmers use functions and libraries for the first time, they frequently refer to textual resources such as code manuals and documentation, to explore and understand the available functionality. Inspired by this observation, we introduce DocPrompting: a natural-language-to-code generation approach that explicitly leverages documentation by (1) retrieving the re

levant documentation pieces given an NL intent, and (2) generating code based on the NL intent and the retrieved documentation. DocPrompting is general: it can be applied to any programming language and is agnostic to the underlying neural model. We demonstrate that DocPrompting consistently improves NL-to-code models: DocPrompting improves strong base models such as CodeT5 by 2.85% in pass@1 (52% relative gain) and 4.39% in pass@10 (30% relative gain) in execution-based evaluation on the popular Python CoNaLa benchmark; on a new Bash dataset tldr, DocPrompting improves CodeT5 and GPT-Neo1.3B by up to absolute 6.9% exact match.

****

Comparing semantic and morphological analogy completion in word embeddings
Haopeng Xie
Word embeddings have prompted great excitement in the NLP community due to their capacity for generalization to unforeseen tasks, including semantic analogy completion. Features such as color and category relationships have been examined by previous work, but this is the first research considering the morphological relationships encoded in word embeddings. We construct several natural experiments examining analogy completion across word stems modified by affixes, and find no evidence that Word2Vec, glove, and fasttext models encode these morphological relationships. We note that a special case of this problem is part-of-speech transformation, and note that the lack of support for part-of-speech analogies is surprising in the context of other successful cases of semantic inference using word embeddings.

****

LipsFormer: Introducing Lipschitz Continuity to Vision Transformers
Xianbiao Qi,Jianan Wang,Yihao Chen,Yukai Shi,Lei Zhang
We present a Lipschitz continuous Transformer, called LipsFormer, to pursue training stability both theoretically and empirically for Transformer-based models. In contrast to previous practical tricks that address training instability by learning rate warmup, layer normalization, attention formulation, and weight initialization, we show that Lipschitz continuity is a more essential property to ensure training stability. In LipsFormer, we replace unstable Transformer component modules with Lipschitz continuous counterparts: CenterNorm instead of LayerNorm, spectral initialization instead of Xavier initialization, scaled cosine similarity attention instead of dot-product attention, and weighted residual shortcut. We prove that these introduced modules are Lipschitz continuous and derive an upper bound on the Lipschitz constant of LipsFormer. Our experiments show that LipsFormer allows stable training of deep Transformer architectures without the need of careful learning rate tuning such as warmup, yielding a faster convergence and better generalization. As a result, on the ImageNet 1K dataset, LipsFormer-Tiny training for 100 epochs without learning rate warmup attains a top-1 accuracy of 81.6\% which is higher than Swin Transformer-Tiny training for 300 epochs with warmup. Moreover, LipsFormer-Tiny training for 300 epochs achieves a top-1 accuracy of 83.5\% with 4.7G FLOPs and 24M parameters.

****

Automatic Chain of Thought Prompting in Large Language Models
Zhuosheng Zhang,Aston Zhang,Mu Li,Alex Smola
Large Language Models (LLMs) can carry out complex reasoning tasks by generating intermediate reasoning steps. These steps are triggered by what is called chain-of-thought (CoT) prompting, which comes in two flavors: one leverages a simple prompt like "Let’s think step by step" to facilitate step-by-step reasoning before answering a question (Zero-Shot-CoT). The other uses manual demonstrations, each composed of a question and a reasoning chain that leads to an answer (Manual-CoT). Unfortunately, the superior performance of the latter strategy crucially hinges on manually generating task-specific demonstrations. This makes it far less scalable and more dependent on the talent of the CoT engineer. We show that such manual efforts may be eliminated by leveraging LLMs to generate the reasoning chains on its own. Since these generated chains often come with mistakes we propose a number of mitigation strategies. Our proposed Auto-CoT method automaticaly samples diverse questions and we perform post-processing quality control to generate usable reasoning chains from Zero-Shot-CoT. On ten public benchmark reas

oning tasks, Auto-CoT performs on par with Manual-CoT without the need for human intervention. Code is available at https://github.com/amazon-research/auto-cot.

****

An efficient encoder-decoder architecture with top-down attention for speech separation

Kai Li,Runxuan Yang,Xiaolin Hu

Deep neural networks have shown excellent prospects in speech separation tasks. However, obtaining good results while keeping a low model complexity remains challenging in real-world applications. In this paper, we provide a bio-inspired efficient encoder-decoder architecture by mimicking the brain’s top-down attention, called TDANet, with decreased model complexity without sacrificing performance. The top-down attention in TDANet is extracted by the global attention (GA) module and the cascaded local attention (LA) layers. The GA module takes multi-scale acoustic features as input to extract global attention signal, which then modulates features of different scales by direct top-down connections. The LA layers use features of adjacent layers as input to extract the local attention signal, which is used to modulate the lateral input in a top-down manner. On three benchmark datasets, TDANet consistently achieved competitive separation performance to previous state-of-the-art (SOTA) methods with higher efficiency. Specifically, TDANet’s multiply-accumulate operations (MACs) are only 5% of Sepformer, one of the previous SOTA models, and CPU inference time is only 10% of Sepformer. In addition, a large-size version of TDANet obtained SOTA results on three datasets, with MACs still only 10% of Sepformer and the CPU inference time only 24% of Sepformer. Our study suggests that top-down attention can be a more efficient strategy for speech separation.

****

Treatment Effect Estimation with Collider Bias and Confounding Bias

Baohong Li,Kun Kuang,Ruoxuan Xiong,Fei Wu

To answer causal questions from observational data, it is important to consider the mechanisms that determine which data values are observed and which are missing. Prior work has considered the treatment assignment mechanism and proposed methods to remove the confounding bias from the common causes of treatment and outcome. However, there are other issues in sample selection, commonly overlooked in prior work, that can bias the treatment effect estimation, such as the issue of censored outcome as a form of collider bias. In this paper, we propose the novel Selection Controlled CounterFactual Regression (SC-CFR) to simultaneously address confounding and collider bias. Specifically, we first calculate the magnitude of the collider bias of different instances by estimating the selection model and then add a control term to remove the collider bias while learning a balanced representation to remove the confounding bias when estimating the outcome model. Our method is shown to provide unbiased treatment effect estimates from observational data with confounding and collider bias. Extensive empirical results on both synthetic and real-world datasets show that our method consistently outperforms benchmarks when both types of biases exist.

****

Adaptive Weight Decay: On The Fly Weight Decay Tuning for Improving Robustness

Amin Ghiasi,Ali Shafahi,Reza Ardekani

We introduce adaptive weight decay, which automatically tunes the hyper-parameter for weight decay during each training iteration. For classification problems, we propose changing the value of the weight-decay hyper-parameter on the fly based on the strength of updates from the classification loss (i.e., gradient of cross-entropy), and the regularization loss (i.e., ℓ_2-norm of the weights). We show that this simple modification can result in large improvements in adversarial robustness — an area which suffers from robust overfitting — without requiring extra data. Specifically, our reformulation results in 20% relative robustness improvement for CIFAR-100, and 10% relative robustness improvement on CIFAR-10 comparing to traditional weight-decay. In addition, this method has other desirable properties, such as less sensitivity to learning rate, and smaller weight norms, which the latter contributes to robustness to overfitting to label noise

, and pruning.
****

Machine Unlearning of Federated Clusters

Chao Pan,Jin Sima,Saurav Prakash,Vishal Rana,Olgica Milenkovic

Federated clustering (FC) is an unsupervised learning problem that arises in a number of practical applications, including personalized recommender and healthcare systems. With the adoption of recent laws ensuring the "right to be forgotten", the problem of machine unlearning for FC methods has become of significant importance. We introduce, for the first time, the problem of machine unlearning for FC, and propose an efficient unlearning mechanism for a customized secure FC framework. Our FC framework utilizes special initialization procedures that we show are well-suited for unlearning. To protect client data privacy, we develop the secure compressed multiset aggregation (SCMA) framework that addresses sparse secure federated learning (FL) problems encountered during clustering as well as more general problems. To simultaneously facilitate low communication complexity and secret sharing protocols, we integrate Reed-Solomon encoding with special evaluation points into our SCMA pipeline, and prove that the client communication cost is logarithmic in the vector dimension. Additionally, to demonstrate the benefits of our unlearning mechanism over complete retraining, we provide a theoretical analysis for the unlearning performance of our approach. Simulation results show that the new FC framework exhibits superior clustering performance compared to previously reported FC baselines when the cluster sizes are highly imbalanced. Compared to completely retraining K-means++ locally and globally for each removal request, our unlearning procedure offers an average speed-up of roughly 84x across seven datasets. Our implementation for the proposed method is available at https://github.com/thupchnsky/mufc.
****

A System for Morphology-Task Generalization via Unified Representation and Behavior Distillation

Hiroki Furuta,Yusuke Iwasawa,Yutaka Matsuo,Shixiang Shane Gu

The rise of generalist large-scale models in natural language and vision has made us expect that a massive data-driven approach could achieve broader generalization in other domains such as continuous control. In this work, we explore a method for learning a single policy that manipulates various forms of agents to solve various tasks by distilling a large amount of proficient behavioral data. In order to align input-output (IO) interface among multiple tasks and diverse agent morphologies while preserving essential 3D geometric relations, we introduce morphology-task graph, which treats observations, actions and goals/task in a unified graph representation. We also develop MxT-Bench for fast large-scale behavior generation, which supports procedural generation of diverse morphology-task combinations with a minimal blueprint and hardware-accelerated simulator. Through efficient representation and architecture selection on MxT-Bench, we find out that a morphology-task graph representation coupled with Transformer architecture improves the multi-task performances compared to other baselines including recent discrete tokenization, and provides better prior knowledge for zero-shot transfer or sample efficiency in downstream multi-task imitation learning. Our work suggests large diverse offline datasets, unified IO representation, and policy representation and architecture selection through supervised learning form a promising approach for studying and advancing morphology-task generalization.
****

Effective Self-Supervised Transformers For Sparse Time Series Data

Alex Labach,Aslesha Pokhrel,Seung Eun Yi,Saba Zuberi,Maksims Volkovs,Rahul G Krishnan

Electronic health records (EHRs) typically contain a wide range of time series data that is characterized by high sparsity and irregular observations. Self-supervised Transformer architectures have shown outstanding performance in a variety of structured tasks in natural language processing and computer vision. However, their use in modelling sparse irregular time series with tabular data has not been widely explored. One of the major challenges is the quadratic scaling of self-attention layers that can significantly limit the input sequence length. In t his work, we introduce TESS, Transformers for EHR data with Self Supervised learning, a self-supervised Transformer-based architecture designed to extract robust representations from EHR data. We propose an input binning scheme that aggregates the time series inputs and sparsity information into a regular sequence with fixed length, enabling the training of larger and deeper Transformers. We demonstrate that significant compression of EHR input data is possible without sacrificing useful information, likely due to the highly correlated nature of observations in small time bins. We then introduce self-supervised prediction tasks that provide rich and informative signals for model pre-training. TESS outperforms state-of-the-art deep learning models on multiple downstream tasks from the MIMIC-IV and PhysioNet-2012 EHR datasets.

****

Scalable feature selection via sparse learnable masks

Yihe Dong,Sercan O Arik

We propose a canonical approach for feature selection, sparse learnable masks (SLM). SLM integrates learnable sparse masks into end-to-end training. For the fundamental non-differentiability challenge of selecting a desired number of features, we propose duo mechanisms for automatic mask scaling to achieve the desired feature sparsity, and gradually tempering this sparsity for effective learning. In addition, SLM employs a novel objective that maximizes the mutual information between the selected features and the labels. Empirically, SLM achieves state-of-the-art results on several benchmark datasets, often by a significant margin, especially on real-world challenging datasets.

****

On the Interplay Between Misspecification and Sub-optimality Gap: From Linear Contextual Bandits to Linear MDPs

Weitong Zhang,Jiafan He,Zhiyuan Fan,Quanquan Gu

We study linear contextual bandits in the misspecified setting, where the expected reward function can be approximated by a linear function class up to a bounded misspecification level $\zeta>0$. We propose an algorithm based on a novel data selection scheme, which only selects the contextual vectors with large uncertainty for online regression. We show that, when the misspecification level ζ is dominated by $\tilde O(\Delta / \sqrt{d})$ with Δ being the minimal sub-optimality gap and d being the dimension of the contextual vectors, our algorithm enjoys the same gap-dependent regret bound $\tilde O ({d^2} /{\Delta})$ as in the well-specified setting up to logarithmic factors. Together with a lower bound adapted from Du et al. (2019); Lattimore et al.(2020), our result suggests an interplay between misspecification level and the sub-optimality gap: (1) the linear contextual bandit model is efficiently learnable when $\zeta \leq \tilde O({\Delta} / \sqrt{d})$; and (2) it is not efficiently learnable when $\zeta \geq \tilde \Omega({\Delta} / {\sqrt{d}})$. We also extend our algorithm to reinforcement learning with linear Markov decision processes (linear MDPs), and obtain a parallel result of gap-dependent regret. Experiments on both synthetic and real-world datasets corroborate our theoretical results.

****

HAS IT REALLY IMPROVED? KNOWLEDGE GRAPH BASED SEPARATION AND FUSION FOR RECOMMENDATION

Ying Tang,Jintian Zhang

In this paper we study the knowledge graph (KG) based recommendation systems. We first design the metric to study the relationship between different SOTA models and find that the current recommendation systems based on knowledge graph have poor ability to retain collaborative filtering signals, and higher-order connectivity would introduce noises. In addition, we explore the collaborative filtering recommendation method using GNN and design the experiment to show that the information learned between GNN models stacked with different layers is different, which provides the explanation for the unstable performance of GNN stacking different layers from a new perspective. According to the above findings, we first design the model-agnostic Cross-Layer Fusion Mechanism without any parameters to improve the performance of GNN. Experimental results on three datasets for collaborative filtering show that Cross-Layer Fusion Mechanism is effective for impro ving GNN performance. Then we design three independent signal extractors to mine the data at three different perspectives and train them separately. Finally, we use the signal fusion mechanism to fuse different signals. Experimental results on three datasets that introduce KG show that our KGSF achieves significant improvements over current SOTA KG based recommendation methods and the results are interpretable.

****

Counterfactual Contrastive Learning for Robust Text Classification

Xiaosong Yuan,Renchu Guan,Wanli Zuo,Yijia Zhang

Text classification has recently been promoted by large pre-trained language models (PLMs) which aim to identify target classes with knowledge transferred from sets of reading comprehension tasks. However, derivative models of PLMs still suffer from sensitive performance on different datasets, the reasons are multiple such as cross-domain and label imbalance problems, from which most models may learn the spurious correlation between texts and labels. Existing research requires people to manually add counterfactual samples to the dataset or automatically match so-called counterfactual pairs that are already in the dataset for augmentation. In this paper, we propose a novel LDA-based counterfactual contrastive learning framework and three data augmentation methods, to capture the causal information in texts, which can promote the robustness of text classification. To confirm the effectiveness of our proposed model and methods, we design and conduct several couples of experiments. Experimental results demonstrate that our model works well on five popular text classification datasets on distinct tasks, we find that training with proposed data augmentation outperforms other augmentation methods on many superior models by 1\% or above. Plus, robustness tests on different datasets also show a competitive performance, which proves the effectiveness of our model and data.

****

SAM as an Optimal Relaxation of Bayes

Thomas Möllenhoff,Mohammad Emtiyaz Khan

Sharpness-aware minimization (SAM) and related adversarial deep-learning methods can drastically improve generalization, but their underlying mechanisms are not yet fully understood. Here, we establish SAM as a relaxation of the Bayes objective where the expected negative-loss is replaced by the optimal convex lower bound, obtained by using the so-called Fenchel biconjugate. The connection enables a new Adam-like extension of SAM to automatically obtain reasonable uncertainty estimates, while sometimes also improving its accuracy. By connecting adversarial and Bayesian methods, our work opens a new path to robustness.

****

Denoising MCMC for Accelerating Diffusion-Based Generative Models

Beomsu Kim,Jong Chul Ye

Diffusion models are powerful generative models that simulate the reverse of diffusion processes using score functions to synthesize data from noise. The sampling process of diffusion models can be interpreted as solving the reverse stochastic differential equation (SDE) or the ordinary differential equation (ODE) of the diffusion process, which often requires up to thousands of discretization steps to generate a single image. This has sparked a great interest in developing efficient integration techniques for reverse-S/ODEs. Here, we propose an orthogonal approach to accelerating score-based sampling: Denoising MCMC (DMCMC). DMCMC first uses MCMC to produce samples in the product space of data and variance (or diffusion time). Then, a reverse-S/ODE integrator is used to denoise the MCMC samples. Since MCMC traverses close to the data manifold, the computation cost of producing a clean sample for DMCMC is much less than that of producing a clean sample from noise. To verify the proposed concept, we show that Denoising Langevin Gibbs (DLG), an instance of DMCMC, successfully accelerates all six reverse-S/ODE integrators considered in this work on the tasks of CIFAR10 and CelebA-HQ-256 image generation. Notably, combined with integrators of Karras et al. (2022) and pre-trained score models of Song et al. (2021b), DLG achieves state-of-the-art results. In the limited number of score function evaluation (NFE) settings on CIFAR10, we have 3.86 FID with ≈ 10 NFE and 2.63 FID with $\approx

20$ NFE. On CelebA-HQ-256, we have 6.99 FID with ≈ 160 NFE, which beats the current best record of Kim et al. (2022) among score-based models, 7.16 FID with 4000 NFE.
****

Learning on Large-scale Text-attributed Graphs via Variational Inference
Jianan Zhao,Meng Qu,Chaozhuo Li,Hao Yan,Qian Liu,Rui Li,Xing Xie,Jian Tang
This paper studies learning on text-attributed graphs (TAGs), where each node is associated with a text description. An ideal solution for such a problem would be integrating both the text and graph structure information with large language models and graph neural networks (GNNs). However, the problem becomes very challenging when graphs are large due to the high computational complexity brought by training large language models and GNNs together. In this paper, we propose an efficient and effective solution to learning on large text-attributed graphs by fusing graph structure and language learning with a variational Expectation-Maximization (EM) framework, called GLEM. Instead of simultaneously training large language models and GNNs on big graphs, GLEM proposes to alternatively update the two modules in the E-step and M-step. Such a procedure allows training the two modules separately while simultaneously allowing the two modules to interact and mutually enhance each other. Extensive experiments on multiple data sets demonstrate the efficiency and effectiveness of the proposed approach.
****

Rethinking Identity in Knowledge Graph Embedding
Jiayi Li,Jiaqi Sun,Yujiu Yang
Knowledge Graph Embedding (KGE) is a common method to complete real-world Knowledge Graphs (KGs) by learning the embeddings of entities and relations.
Beyond specific KGE models, previous work proposes a general framework based on group. A group has a special element identity that uniquely corresponds to the relation identity in KGs, which implies that identity should be represented uniquely. However, we find that this uniqueness cannot be modeled by bilinear based models, revealing the crack between the framework and models. To this end, we study the required conditions and propose a solution named Unit Ball Bilinear Model (UniBi). In addition to the theoretical superiority, UniBi is more robust and interpretable. Experiments demonstrate that UniBi models the uniqueness without the cost of performance and verify its a robustness and interpretability.
****

Eigen Memory Trees
Mark Rucker,Jordan T. Ash,John Langford,Paul Mineiro,Ida Momennejad
This work introduces the Eigen Memory Tree (EMT), a novel online memory model for sequential learning scenarios. EMTs store data at the leaves of a binary tree, and route new samples through the structure using the principal components of previous experiences, facilitating efficient (logarithmic) access to relevant memories. We demonstrate that EMT outperforms existing online memory approaches, and provide a hybridized EMT-parametric algorithm that enjoys drastically improved performance over purely parametric methods with nearly no downsides. Our findings are validated using 206 datasets from OpenML repository in both bounded and infinite memory budget situations.
****

Energy-based Predictive Representation for Reinforcement Learning
Tianjun Zhang,Tongzheng Ren,Chenjun Xiao,Wenli Xiao,Joseph E. Gonzalez,Dale Schuurmans,Bo Dai
In real world applications, it is usually necessary for a reinforcement learning algorithm to handle the partial observability beyond Markov decision processes (MDPs). Although the partially observable Markov decision process (POMDP) has been precisely motivated for this requirement, such a formulation raises significant computational and statistical hardness challenges in learning and planning. In this work, we introduce the Energy-based Predictive Representation (EPR), which leads to a unified framework for practical reinforcement learning algorithm design in both MDPs and POMDPs settings, to handle the learning, exploration, and planning in a coherent way. The proposed approach relies on the powerful neural energy-based model to extract sufficient representation, from which Q-functions

can be efficiently approximated. With such a representation, we develop an efficient approach for computing confidence, which allows optimism/pessimism in the face of uncertainty to be efficiently implemented in planning, hence managing the exploration versus exploitation tradeoff. An experimental investigation shows that the proposed algorithm can surpass state-of-the-art performance in both MDP and POMDP settings in comparison to existing baselines.

****

Which Invariance Should We Transfer? A Causal Minimax Learning Approach

Mingzhou Liu,Xiangyu Zheng,Xinwei Sun,Fang Fang,Yizhou Wang

A major barrier to deploy current machine learning models lies in their sensitivity to dataset shifts. To resolve this problem, most existing studies attempted to transfer stable information to unseen environments. Among these, graph-based methods causally decomposed the data generating process into stable and mutable mechanisms. By removing the effect of mutable generation, they identified a set of stable predictors. However, a key question regarding robustness remains: which subset of the whole stable information should the model transfer, in order to achieve optimal generalization ability? To answer this question, we provide a comprehensive minimax analysis that fully characterizes conditions for a subset to be optimal. Particularly in general cases, we propose to maximize over mutable mechanisms (i.e., the source of dataset shifts), which is provable to identify the worst-case risk over all environments. This ensures us to select the optimal subset with the minimal worst-case risk. To reduce computational costs, we propose to search over only equivalent classes in terms of worst-case risk, instead of over all subsets. In cases when the searching space is still large, we turn this subset selection problem into a sparse min-max optimization scheme, which enjoys the simplicity and efficiency of implementation. The utility of our methods is demonstrated on the diagnosis of Alzheimer's Disease and gene function prediction.

****

Exclusive Supermask Subnetwork Training for Continual Learning

Prateek Yadav,Mohit Bansal

Continual Learning (CL) methods mainly focus on avoiding catastrophic forgetting and learning representations that are transferable to new tasks. Recently, Wortsman et al. (2020) proposed a CL method, SupSup, which uses a randomly initialized, fixed base network (model) and finds a supermask for each new task that selectively keeps or removes each weight to produce a subnetwork. They prevent forgetting as the network weights are not being updated. Although there is no forgetting, the performance of supermask is sub-optimal because fixed weights restrict its representational power. Furthermore, there is no accumulation or transfer of knowledge inside the model when new tasks are learned. Hence, we propose ExSSNeT (Exclusive Supermask SubNEtwork Training), that performs exclusive and non-overlapping subnetwork weight training. This avoids conflicting updates to the shared weights by subsequent tasks to improve performance while still preventing forgetting. Furthermore, we propose a novel KNN-based Knowledge Transfer (KKT) module that dynamically initializes a new task's mask based on previous tasks for improving knowledge transfer. We demonstrate that ExSSNeT outperforms SupSup and other strong previous methods on both text classification and vision tasks while preventing forgetting. Moreover, ExSSNeT is particularly advantageous for sparse masks that activate 2-10% of the model parameters, resulting in an average improvement of 8.3% over SupSup. Additionally, ExSSNeT scales to a large number of tasks (100), and our KKT module helps to learn new tasks faster while improving the overall performance.

****

Dual personalization for federated recommendation on devices

Chunxu Zhang,Guodong Long,Tianyi Zhou,Zijian Zhang,Bo Yang

Federated recommendation is a new Internet service architecture that aims to provide privacy-preserving recommendation services in federated settings. Existing solutions are used to combine distributed recommendation algorithms and privacy-preserving mechanisms. Thus it inherently takes the form of heavyweight models at the server and hinders the deployment of on-device intelligent models to end-u

sers. This paper proposes a novel Personalized Federated Recommendation (PFedRec) framework to learn many user-specific lightweight models to be deployed on smart devices rather than a heavyweight model on a server. Moreover, we propose a new dual personalization mechanism to effectively learn fine-grained personalization on both users and items. The overall learning process is formulated into a unified federated optimization framework. Specifically, unlike previous methods that share exactly the same item embeddings across users in a federated system, dual personalization allows mild finetuning of item embeddings for each user to generate user-specific views for item representations which can be integrated into existing federated recommendation methods to gain improvements immediately. Experiments on multiple benchmark datasets have demonstrated the effectiveness of PFedRec and the dual personalization mechanism. Moreover, we provide visualizations and in-depth analysis of the personalization techniques in item embedding, which shed novel insights on the design of RecSys in federated settings.
****

Moderate Coreset: A Universal Method of Data Selection for Real-world Data-efficient Deep Learning

Xiaobo Xia,Jiale Liu,Jun Yu,Xu Shen,Bo Han,Tongliang Liu

Deep learning methods nowadays rely on massive data, resulting in substantial costs of data storage and model training. Data selection is a useful tool to alleviate such costs, where a coreset of massive data is extracted to practically perform on par with full data. Based on carefully-designed score criteria, existing methods first count the score of each data point and then select the data points whose scores lie in a certain range to construct a coreset. These methods work well in their respective preconceived scenarios but are not robust to the change of scenarios, since the optimal range of scores varies as the scenario changes. The issue limits the application of these methods, because realistic scenarios often mismatch preconceived ones, and it is inconvenient or unfeasible to tune the criteria and methods accordingly. In this paper, to address the issue, a concept of the moderate coreset is discussed. Specifically, given any score criterion of data selection, different scenarios prefer data points with scores in different intervals. As the score median is a proxy of the score distribution in statistics, the data points with scores close to the score median can be seen as a proxy of full data and generalize different scenarios, which are used to construct the moderate coreset. As a proof-of-concept, a universal method that inherits the moderate coreset and uses the distance of a data point to its class center as the score criterion, is proposed to meet complex realistic scenarios. Extensive experiments confirm the advance of our method over prior state-of-the-art methods, leading to a strong baseline for future research. The implementation is available at https://github.com/tmllab/Moderate-DS.
****

Effectively Clarify Confusion via Visualized Aggregation and Separation of Deep Representation

Yi Lin,Jingchi Jiang,Dongxin Chen,Zhaoyang Ma,Yi Guan,Xiguang Liu,Haiyan You,Jing Yang

Clarifying confusion is the most critical issue for improving classification performance. Confusion occurs with almost all classification models but tends to be ignored in excellent-performing models. The current mainstream research mainly focuses on solving the confusion in a specific case, such as data insufficiency and class imbalance. We believe that mining the commonalities of the same class and the gaps among different classes will effectively clarify the confusion. In this paper, we propose a novel, simple and intuitive Aggregation Separation Loss (ASLoss), as an adjunct for classification loss to clarify the confusion in some common cases. The ASLoss aggregates the representations of the same class samples as near as possible and separates the representations of different classes as far as possible.

We use two image classification tasks with three simultaneous confounding characteristics i.e. data insufficiency, class imbalance, and unclear class evidence to demonstrate ASLoss. Representation visualization, confusion comparison and detailed comparison experiments are conducted. The results show that representation s in deep spaces extracted by ASLoss are sufficiently clear and distinguishable, the confusion among different classes is significantly clarified and the optimal network using ASLoss reaches the state-of-the-art level.

****

Time-Transformer AAE: Connecting Temporal Convolutional Networks and Transformer for Time Series Generation

Yuansan Liu,Sudanthi Wijewickrema,Ang Li,James Bailey

Generating time series data is a challenging task due to the complex temporal properties of this type of data. Such temporal properties typically include local correlations as well as global dependencies. Most existing generative models have failed to effectively learn both the local and global properties of time series data. To address this open problem, we propose a novel time series generative model consisting of an adversarial autoencoder (AAE) and a newly designed architecture named `Time-Transformer' within the decoder. We call this generative model `Time-Transformer AAE'. The Time-Transformer first simultaneously learns local and global features in a layer-wise parallel design, combining the abilities of Temporal Convolutional Networks (TCN) and Transformer in extracting local features and global dependencies respectively. Second, a bidirectional cross attention is proposed to provide complementary guidance across the two branches and achieve proper fusion between local and global features. Experimental results demonstrate that our model can outperform existing state-of-the-art models in most cases, especially when the data contains both global and local properties. We also show our model's ability to perform a downstream task: data augmentation to support the solution of imbalanced classification problems.

****

A comparison of dataset distillation and active learning in text classification

Zhang Zeyu,Zhang Yuanchun

Deep learning has achieved great success over the past few years in different aspects ranging from computer vision to natural language process. However, the huge size of data in deep learning has always been a thorny problem in learning the underlying distribution and tackling various human tasks. To alleviate this problem, knowledge distillation has been proposed to simplify the model, and later dataset distillation as a new method of reducing dataset sizes has been proposed, which aims to synthesize a small number of samples that contain all the information of a very large dataset. Meanwhile, active learning is also an effective method to reduce dataset sizes by only selecting the most significant labeling samples from the original dataset. In this paper, we explore the discrepancies in the principles of dataset distillation and active learning, and evaluate two algorithms on NLP classification dataset: Stanford Sentiment Treebank. The result of the experiment is that the distilled data with the size of 0.1% of the original text data achieves approximately 88% accuracy, while the selected data achieves 52% performance of the original data.

****

Temporally Consistent Video Transformer for Long-Term Video Prediction

Wilson Yan,Danijar Hafner,Stephen James,Pieter Abbeel

Generating long, temporally consistent video remains an open challenge in video generation. Primarily due to computational limitations, most prior methods limit themselves to training on a small subset of frames that are then extended to generate longer videos through a sliding window fashion. Although these techniques may produce sharp videos, they have difficulty retaining long-term temporal consistency due to their limited context length. In this work, we present Temporally Consistent Video Transformer (TECO), a vector-quantized latent dynamics video prediction model that learns compressed representations to efficiently condition on long videos of hundreds of frames during both training and generation. We use a MaskGit prior for dynamics prediction which enables both sharper and faster generations compared to prior work. Our experiments show that TECO outperforms SOTA baselines in a variety of video prediction benchmarks ranging from simple mazes in DMLab, large 3D worlds in Minecraft, and complex real-world videos from Kinetics-600. In addition, to better understand the capabilities of video predict ion models in modeling temporal consistency, we introduce several challenging video prediction tasks consisting of agents randomly traversing 3D scenes of varying difficulty. This presents a challenging benchmark for video prediction in partially observable environments where a model must understand what parts of the scenes to re-create versus invent depending on its past observations or generations. An anonymized website with samples can be found at https://sites.google.com/view/iclr23-teco

****

Extreme Q-Learning: MaxEnt RL without Entropy

Divyansh Garg,Joey Hejna,Matthieu Geist,Stefano Ermon

Modern Deep Reinforcement Learning (RL) algorithms require estimates of the maximal Q-value, which are difficult to compute in continuous domains with an infinite number of possible actions. In this work, we introduce a new update rule for online and offline RL which directly models the maximal value using Extreme Value Theory (EVT), drawing inspiration from economics. By doing so, we avoid computing Q-values using out-of-distribution actions which is often a substantial source of error. Our key insight is to introduce an objective that directly estimates the optimal soft-value functions (LogSumExp) in the maximum entropy RL setting without needing to sample from a policy. Using EVT, we derive our \emph{Extreme Q-Learning} framework and consequently online and, for the first time, offline MaxEnt Q-learning algorithms, that do not explicitly require access to a policy or its entropy. Our method obtains consistently strong performance in the D4RL benchmark, outperforming prior works by \emph{10+ points} on the challenging Franka Kitchen tasks while offering moderate improvements over SAC and TD3 on online DM Control tasks. Visualizations and code can be found on our website.

****

Autoencoding Hyperbolic Representation for Adversarial Generation

Eric Qu,Dongmian Zou

With the recent advance of geometric deep learning, neural networks have been extensively used for data in non-Euclidean domains. In particular, hyperbolic neural networks have proved successful in processing hierarchical information of data. However, many hyperbolic neural networks are numerically unstable during training, which precludes using complex architectures. This crucial problem makes it difficult to build hyperbolic generative models for real and complex data. In this work, we propose a hyperbolic generative network in which we design novel architecture and layers to guarantee stable training. Our proposed network contains three parts: first, a hyperbolic autoencoder (AE) that produces hyperbolic embedding for input data; second, a hyperbolic generative adversarial network (GAN) for generating the hyperbolic latent embedding of the AE from simple noise; third, a generator that inherits the decoder from the AE and the generator from the GAN. We call this network the hyperbolic AE-GAN, or HAEGAN for short. The architecture of HAEGAN fosters expressive representation in the hyperbolic space, and the specific design of layers ensures numerical stability. Experiments show that HAEGAN is able to generate complex data with state-of-the-art structure-related performance.

****

CAREER: Transfer Learning for Economic Prediction of Labor Data

Keyon Vafa,Emil Palikot,Tianyu Du,Ayush Kanodia,Susan Athey,David Blei

Labor economists regularly analyze employment data by fitting predictive models to small, carefully constructed longitudinal survey datasets. Although modern machine learning methods offer promise for such problems, these survey datasets are too small to take advantage of them. In recent years large datasets of online resumes have also become available, providing data about the career trajectories of millions of individuals. However, standard econometric models cannot take advantage of their scale or incorporate them into the analysis of survey data. To this end we develop CAREER, a transformer-based model that uses transfer learning to learn representations of job sequences. CAREER is first fit to large, passively-collected resume data and then fine-tuned to smaller, better-curated datasets for economic inferences. We fit CAREER to a dataset of 24 million job sequences from resumes, and fine-tune its representations on longitudinal survey datas ets. We find that CAREER forms accurate predictions of job sequences, achieving state-of-the-art predictive performance on three widely-used economics datasets. We further find that CAREER can be used to form good predictions of other downstream variables; incorporating CAREER into a wage model provides better predictions than the econometric models currently in use.

****

Federated Nearest Neighbor Machine Translation

Yichao Du,Zhirui Zhang,Bingzhe Wu,Lemao Liu,Tong Xu,Enhong Chen

To protect user privacy and meet legal regulations, federated learning (FL) is attracting significant attention. Training neural machine translation (NMT) models with traditional FL algorithm (e.g., FedAvg) typically relies on multi-round model-based interactions. However, it is impractical and inefficient for machine translation tasks due to the vast communication overheads and heavy synchronization. In this paper, we propose a novel federated nearest neighbor (FedNN) machine translation framework that, instead of multi-round model-based interactions, leverages one-round memorization-based interaction to share knowledge across different clients to build low-overhead privacy-preserving systems. The whole approach equips the public NMT model trained on large-scale accessible data with a $k$-nearest-neighbor ($k$NN) classifier and integrates the external datastore constructed by private text data in all clients to form the final FL model. A two-phase datastore encryption strategy is introduced to achieve privacy-preserving during this process. Extensive experiments show that FedNN significantly reduces computational and communication costs compared with FedAvg, while maintaining promising performance in different FL settings.

****

Latent Variable Representation for Reinforcement Learning

Tongzheng Ren,Chenjun Xiao,Tianjun Zhang,Na Li,Zhaoran Wang,sujay sanghavi,Dale Schuurmans,Bo Dai

Deep latent variable models have achieved significant empirical successes in model-based reinforcement learning (RL) due to their expressiveness in modeling complex transition dynamics. On the other hand, it remains unclear theoretically and empirically how latent variable models may facilitate learning, planning, and exploration to improve the sample efficiency of RL. In this paper, we provide a representation view of the latent variable models for state-action value functions, which allows both tractable variational learning algorithm and effective implementation of the optimism/pessimism principle in the face of uncertainty for exploration. In particular, we propose a computationally efficient planning algorithm with UCB exploration by incorporating kernel embeddings of latent variable models. Theoretically, we establish the sample complexity of the proposed approach in the online and offline settings. Empirically, we demonstrate superior performance over current state-of-the-art algorithms across various benchmarks.

****

Look in The Mirror: Molecular Graph Contrastive Learning with Line Graph

Junran Wu,Xueyuan Chen,Bowen Shi,Jiaheng Liu,Shangzhe Li,Ke Xu

Trapped by the label scarcity in molecular property prediction and drug design, graph contrastive learning came forward. A general contrastive model consists of a view generator, view encoder, and contrastive loss, in which the view mainly controls the encoded information underlying input graphs. Leading contrastive learning works show two kinds of view generators, that is, random or learnable data corruption and domain knowledge incorporation. While effective, the two ways also lead to molecular semantics altering and limited generalization capability, respectively. Thus, a decent view that can fully retain molecular semantics and is free from profound domain knowledge is supposed to come forward. To this end, we relate molecular graph contrastive learning with the line graph and propose a novel method termed LGCL. Specifically, by contrasting the given graph with the corresponding line graph, the graph encoder can freely encode the molecular semantics without omission. While considering the information inconsistency and over-smoothing derived from the learning process because of the mismatched pace of message passing in two kinds of graphs, we present a new patch with edge attribute fusion and two local contrastive losses for performance fixing. Compared wit h state-of-the-art (SOTA) methods for view generation, superior performance on molecular property prediction suggests the effectiveness of line graphs severing as the contrasting views.

****

Precision Collaboration for Federated Learning

Sen Cui,Abudukelimu Wuerkaixi,Jian Liang,Weishen Pan,Jianwei Zhang,Changshui Zhang,Fei Wang

Inherent heterogeneity of local data distributions, which causes inefficient model learning and significant degradation of model performance, has been a key challenge in Federated Learning (FL). So far, plenty of efforts have focused on addressing data heterogeneity by relying on a hypothetical clustering structure or a consistent information sharing mechanism. However, because of the diversity of the real-world local data, these assumptions may be largely violated. In this work, we argue that information sharing is mostly fragmented in the federated network in reality. More specifically, the distribution overlaps are not consistent but scattered in local clients. In this work, we propose the concept ``Precision Collaboration'' which refers to learning from the informative overlaps precisely while avoiding the potential negative transfer induced by others. In particular, we propose to infer the local data manifolds and estimate the exact local data density simultaneously. The learned manifold aims to precisely identify the overlaps from other clients, and the estimated likelihood allows to generate samples from the manifold in an optimal sampling density. Experiments show that our proposed PCFL significantly overcomes baselines on benchmarks and a real-world clinical scenario.

****

RLSBench: A Large-Scale Empirical Study of Domain Adaptation Under Relaxed Label Shift

Saurabh Garg,Nick Erickson,James Sharpnack,Alex Smola,Sivaraman Balakrishnan,Zachary Chase Lipton

Despite the emergence of principled methods for domain adaptation under label shift (where only the class balance changes), the sensitivity of these methods to natural-seeming covariate shifts remains precariously underexplored. Meanwhile, popular deep domain adaptation heuristics, despite showing promise on benchmark datasets, tend to falter when faced with shifts in the class balance. Moreover, it's difficult to assess the state of the field owing to inconsistencies among relevant papers in evaluation criteria, datasets, and baselines. In this paper, we introduce \textsc{RLSbench}, a large-scale benchmark for such \emph{relaxed label shift} settings, consisting of 11 vision datasets spanning $>$200 distribution shift pairs with different class proportions. We evaluate 12 popular domain adaptation methods, demonstrating a more widespread susceptibility to failure under extreme shifts in the class proportions than was previously known. We develop an effective meta-algorithm, compatible with most deep domain adaptation heuristics, that consists of the following two steps: (i) \emph{pseudo-balance} the data at each epoch; and (ii) adjust the final classifier with (an estimate of) target label distribution. Furthermore, we discover that batch-norm adaption of a model trained on source with aforementioned corrections offers a strong baseline, largely missing from prior comparisons. We hope that these findings and the availability of \textsc{RLSbench} will encourage researchers to include rigorously evaluate proposed methods in relaxed label shift settings. Code is publicly available at https://github.com/ICLR2023Anon.

****

ROCO: A General Framework for Evaluating Robustness of Combinatorial Optimization Solvers on Graphs

Han Lu,Zenan Li,Runzhong Wang,Qibing Ren,Xijun Li,Mingxuan Yuan,Jia Zeng,Xiaokang Yang,Junchi Yan

Solving combinatorial optimization (CO) on graphs has been attracting increasing interests from the machine learning community whereby data-driven approaches were recently devised to go beyond traditional manually-designated algorithms. In this paper, we study the robustness of a combinatorial solver as a blackbox regardless it is classic or learning-based though the latter can often be more interesting to the ML community. Specifically, we develop a practically feasible robustness metric for general CO solvers. A no-worse optimal cost guarantee is developed as such the optimal solutions are not required to achieve for solvers, and we tackle the non-differentiable challenge in input instance disturbance by resorting to black-box adversarial attack methods. Extensive experiments are conducted on 14 unique combinations of solvers and CO problems, and we demonstrate that the performance of state-of-the-art solvers like Gurobi can degenerate by over 20% under the given time limit bound on the hard instances discovered by our robustness metric, raising concerns about the robustness of combinatorial optimization solvers.

****

Spatial reasoning as Object Graph Energy Minimization

Nikolaos Gkanatsios,Ayush Jain,Zhou Xian,Yunchu Zhang,Katerina Fragkiadaki

We propose a model that maps spatial rearrangement instructions to goal scene configurations via gradient descent on a set of relational energy functions over object 2D overhead locations, one per spatial predicate in the instruction. Energy based models over object locations are trained from a handful of examples of object arrangements annotated with the corresponding spatial predicates. Predicates can be binary (e.g., left of, right of, etc.) or multi-ary (e.g., circles, lines, etc.). A language parser maps language instructions to the corresponding set of EBMs, and a visual-language model grounds their arguments on relevant objects in the visual scene. Energy minimization on the joint set of energies iteratively updates the object locations till their final configuration. Then, low-level local policies re-locate objects to the inferred goal locations. Our framework shows many forms of strong generalization: (i)joint energy minimization handles zero-shot complex predicate compositions while each EBM is trained only from single predicate instructions, (ii) the model can execute instructions zero-shot, without a need for paired instruction-action training, (iii) instructions can mention novel objects and attributes at test time thanks to the pre-training of the visual language grounding model from large scale passive captioned datasets. We test the model in established instruction-guided manipulation benchmarks, as well as a benchmark of compositional instructions we introduce in this work. We show large improvements over state-of-the-art end-to-end language to action policies and planning in large language models, especially for long instructions and multi-ary spatial concepts.

****

Words are all you need? Language as an approximation for human similarity judgments

Raja Marjieh,Pol Van Rijn,Ilia Sucholutsky,Theodore Sumers,Harin Lee,Thomas L. Griffiths,Nori Jacoby

Human similarity judgments are a powerful supervision signal for machine learning applications based on techniques such as contrastive learning, information retrieval, and model alignment, but classical methods for collecting human similarity judgments are too expensive to be used at scale. Recent methods propose using pre-trained deep neural networks (DNNs) to approximate human similarity, but pre-trained DNNs may not be available for certain domains (e.g., medical images, low-resource languages) and their performance in approximating human similarity has not been extensively tested. We conducted an evaluation of 611 pre-trained models across three domains -- images, audio, video -- and found that there is a large gap in performance between human similarity judgments and pre-trained DNNs. To address this gap, we propose a new class of similarity approximation methods based on language. To collect the language data required by these new methods, we also developed and validated a novel adaptive tag collection pipeline. We find that our proposed language-based methods are significantly cheaper, in the number of human judgments, than classical methods, but still improve performance over the DNN-based methods. Finally, we also develop `stacked' methods that combine language embeddings with DNN embeddings, and find that these consistently provide the best approximations for human similarity across all three of our modalit ies. Based on the results of this comprehensive study, we provide a concise guide for researchers interested in collecting or approximating human similarity data. To accompany this guide, we also release all of the similarity and language data, a total of 206,339 human judgments, that we collected in our experiments, along with a detailed breakdown of all modeling results.

****

Graph Contrastive Learning with Reinforced Augmentation
Ziyang Liu,Chaokun Wang

Graph contrastive learning (GCL), designing contrastive objectives to learn embeddings from augmented graphs, has become a prevailing method for learning embeddings from graphs in an unsupervised manner. As an important procedure in GCL, graph data augmentation (GDA) directly affects the model performance on the downstream task. Currently, there are three types of GDA strategies: trial-and-error, precomputed method, and adversarial method. However, these strategies ignore the connection between the two consecutive augmentation results because GDA is regarded as an independent process. In this paper, we regard the GDA in GCL as a Markov decision process. Based on this point, we propose a reinforced method, i.e., the fourth type of GDA strategy, using a novel Graph Advantage Actor-Critic (GA2C) model for GCL. On 23 graph datasets, the experimental results verify that GA2C outperforms the SOTA GCL models on a series of downstream tasks such as graph classification, node classification, and link prediction.

****

A Novel Fast Exact Subproblem Solver for Stochastic Quasi-Newton Cubic Regularized Optimization
Jarad Forristal,Joshua Griffin,Wenwen Zhou,Seyedalireza Yektamaram

In this work we describe an Adaptive Regularization using Cubics (ARC) method for large-scale nonconvex unconstrained optimization using Limited memory Quasi-Newton (LQN) matrices. ARC methods are a relatively new family of second-order optimization strategies that utilize a cubic-regularization (CR) term in place of trust-regions or line-searches. Solving the CR subproblem exactly requires Newton's method, yet using properties of the internal structure of LQN matrices, we are able to find exact solutions to the CR subproblem in a matrix-free manner, providing very large speedups. Additionally, we expand upon previous ARC work and explicitly incorporate first-order updates into our algorithm. We provide empirical results for different LQN matrices and find our proposed method compares to or exceeds all tested optimizers with minimal tuning.

****

Block-Diagonal Structure Learning for Subspace Clustering
Zheng Xing,Weibing Zhao

Finding the informative subspaces of high-dimensional ordered datasets is at the core of innumerable applications in computer vision, where spectral-based subspace clustering is arguably the most commonly studied method due to its strong empirical performance. Such algorithms compute an affinity matrix to construct a self-representation for each sample utilizing other samples as a dictionary, and spectral clustering is employed to identify the clustering structure based on the affinity matrix. Since the ordered nature, the block-diagonal structure learning embedded in self-representation plays a vital role in effective subspace clustering. However, direct optimization with block-diagonal priors is challenging due to the random sparseness and connectivity nature of self-representation, and none of the existing techniques resort to the block-diagonal structure learning of self-representation alone. In this paper, we propose a technique, namely block-diagonal structure representation learning, to solve the optimal clustering of the ordered data directly instead of employing spectral clustering. The proposed algorithm can theoretically achieve the global optimal solution of the proposed discrete non-convex block-diagonal partition problem. We test the proposed clustering method on several types of segmentation databases, such as human face recognization, video scene clip partition, motion tracks, and dynamic 3-D facial expression sequences. The experiments illustrate that the proposed method outperforms state-of-the-art subspace clustering methods.

****

Offline RL of the Underlying MDP from Heterogeneous Data Sources
Chengshuai Shi,Wei Xiong,Cong Shen,Jing Yang

Most of the existing offline reinforcement learning (RL) studies assume the available dataset is sampled directly from the target environment. However, in some practical applications, the available data are often coming from several related but heterogeneous environments. A theoretical understanding of efficient learning from heterogeneous offline datasets remains lacking. In this work, we study the problem of learning a (hidden) underlying Markov decision process (MDP) based on heterogeneous offline datasets collected from multiple randomly perturbed data sources. A novel HetPEVI algorithm is proposed, which jointly considers two types of uncertainties: sample uncertainties from the finite number of data samples per data source, and source uncertainties due to a finite number of data sources. Building on HetPEVI, we further incorporate reference-advantage decompositions and Bernstein-type penalties to propose the HetPEVI-Adv algorithm. Theoretical analysis not only proves the effectiveness of both HetPEVI and HetPEVI-Adv but also demonstrates the advantage of the latter. More importantly, the results explicitly characterize the learning loss due to the finite heterogeneously realized environments compared with sampling directly from the underlying MDP. Finally, we extend the study to MDPs with linear function approximation and propose the HetPEVI-Lin algorithm that provides additional efficiency guarantees beyond the tabular case.

****

FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning
Yidong Wang,Hao Chen,Qiang Heng,Wenxin Hou,Yue Fan,Zhen Wu,Jindong Wang,Marios Savvides,Takahiro Shinozaki,Bhiksha Raj,Bernt Schiele,Xing Xie

Semi-supervised Learning (SSL) has witnessed great success owing to the impressive performances brought by various methods based on pseudo labeling and consistency regularization. However, we argue that existing methods might fail to utilize the unlabeled data more effectively since they either use a pre-defined / fixed threshold or an ad-hoc threshold adjusting scheme, resulting in inferior performance and slow convergence. We first analyze a motivating example to obtain intuitions on the relationship between the desirable threshold and model's learning status. Based on the analysis, we hence propose FreeMatch to adjust the confidence threshold in a self-adaptive manner according to the model's learning status. We further introduce a self-adaptive class fairness regularization penalty to encourage the model for diverse predictions during the early training stage. Extensive experiments indicate the superiority of FreeMatch especially when the labeled data are extremely rare. FreeMatch achieves 5.78%, 13.59%, and 1.28% error rate reduction over the latest state-of-the-art method FlexMatch on CIFAR-10 with 1 label per class, STL-10 with 4 labels per class, and ImageNet with 100 labels per class, respectively. Moreover, FreeMatch can also boost the performance of imbalanced SSL. The codes can be found at https://github.com/microsoft/Semi-supervised-learning.

****

Training image classifiers using Semi-Weak Label Data
Ankit Parag Shah,Bhiksha Raj

This paper introduces a new semi-weak label learning paradigm which provides additional information in comparison to the weak label classification. We define semi-weak label data as data where we know the presence or absence of a given class and additionally we have the information about the exact count of each class as opposed to knowing the label proportions. A three-stage framework is proposed to address the problem of learning from semi-weak labels. It leverages the fact that counting information is naturally non-negative and discrete. Experiments are conducted on generated samples from CIFAR-10 and we compare our model with a fully-supervised setting baseline, a weakly-supervised setting baseline and a learning from proportion(LLP) baseline. Our framework not only outperforms both baseline models for MIL-based weakly supervised setting and learning from proportio n setting, but also gives comparable results compared to the fully supervised model. Further, we conduct thorough ablation studies to analyze across datasets and variation with batch size, losses architectural changes, bag size and regularization, thereby demonstrating robustness of our approach.

****

Confidence Estimation Using Unlabeled Data

Chen Li,Xiaoling Hu,Chao Chen

Overconfidence is a common issue for deep neural networks, limiting their deployment in real-world applications. To better estimate confidence, existing methods mostly focus on fully-supervised scenarios and rely on training labels. In this paper, we propose the first confidence estimation method for a semi-supervised setting, when most training labels are unavailable. We stipulate that even with limited training labels, we can still reasonably approximate the confidence of model on unlabeled samples by inspecting the prediction consistency through the training process. We use training consistency as a surrogate function and propose a consistency ranking loss for confidence estimation. On both image classification and segmentation tasks, our method achieves state-of-the-art performances in confidence estimation. Furthermore, we show the benefit of the proposed method through a downstream active learning task.

****

Spectral Decomposition Representation for Reinforcement Learning

Tongzheng Ren,Tianjun Zhang,Lisa Lee,Joseph E. Gonzalez,Dale Schuurmans,Bo Dai

Representation learning often plays a critical role in avoiding the curse of dimensionality in reinforcement learning. A representative class of algorithms exploits spectral decomposition of the stochastic transition dynamics to construct representations that enjoy strong theoretical properties in idealized settings. However, current spectral methods suffer from limited applicability because they are constructed for
state-only aggregation and are derived from a policy-dependent transition kernel, without considering the issue of exploration. To address these issues, we propose an alternative spectral method, Spectral Decomposition Representation (SPEDER), that extracts a state-action abstraction from the dynamics without inducing spurious dependence on the data collection policy, while also balancing the exploration-versus-exploitation trade-off during learning. A theoretical analysis establishes the sample efficiency of the proposed algorithm in both the online and offline settings. In addition, an experimental investigation demonstrates superior performance over current state-of-the-art algorithms across several RL benchmarks.

****

On Accelerated Perceptrons and Beyond

Guanghui Wang,Rafael Hanashiro,Etash Kumar Guha,Jacob Abernethy

The classical Perceptron algorithm of Rosenblatt can be used to find a linear threshold function to correctly classify n linearly separable data points, assuming the classes are separated by some margin $\gamma > 0$. A foundational result is that Perceptron converges after $\Omega(1/\gamma^{2})$ iterations. There have been several recent works that managed to improve this rate by a quadratic factor, to $\Omega(\sqrt{\log n}/\gamma)$, with more sophisticated algorithms. In this paper, we unify these existing results under one framework by showing that they can all be described through the lens of solving min-max problems using modern acceleration techniques, mainly through \emph{optimistic} online learning.
We then show that the proposed framework also leads to improved results for a series of problems beyond the standard Perceptron setting. Specifically, a) for the margin maximization problem, we improve the state-of-the-art result from $O(\log t/t^2)$ to $O(1/t^2)$, where t is the number of iterations; b) we provide the first result on identifying the implicit bias property of the classical Nesterov's accelerated gradient descent (NAG) algorithm, and show NAG can maximize the margin with an $O(1/t^2)$ rate; c) for the classical p-norm Perceptron problem, we provide an algorithm with $\Omega(\sqrt{(p-1)\log n}/\gamma)$ convergence rate, while existing algorithms suffer the $\Omega({(p-1)}/\gamma^2)$ convergence rate.

****

DITTO: Offline Imitation Learning with World Models

Branton DeMoss,Paul Duckworth,Nick Hawes,Ingmar Posner

We propose DITTO, a fully offline approach to imitation learning which addresses the problem of covariate shift without access to an oracle or any additional online interactions. By unrolling agent policies in the latent space of a learned world model and penalizing drift from expert demonstrations, we can use online reinforcement learning algorithms to learn policies which solve the imitation objective, without access to the underlying environment or reward function. Decoupling policy and world model learning lets us leverage datasets of any quality to learn latent representations which provide a natural reward signal for imitation learning, avoiding the need for complex adversarial or sparse imitation-inducing rewards. Compared to competitive baselines, our method achieves state-of-the-art performance in a variety of challenging environments from pixel observations alone.

****

BAT-Chain: Bayesian-Aware Transport Chain for Topic Hierarchies Discovery

Dongsheng Wang,He Zhao,Dan Dan Guo,Xinyang Liu,Miaoge Li,Bo Chen,Mingyuan Zhou

Topic modeling has been an important tool for text analysis. Originally, topics discovered by a model are usually assumed to be independent. However, as a semantic representation of a concept, a topic is naturally related to others, which motivates the development of learning hierarchical topic structure. Most existing Bayesian models are designed to learn hierarchical structure, but they need non-trivial posterior inference. Although the recent transport-based topic models bypass the posterior inference, none of them considers deep topic structures. In this paper, we interpret document as its word embeddings and propose a novel Bayesian-aware transport chain to discover multi-level topic structures, where each layer learns a set of topic embeddings and the document hierarchical representations are defined as a series of empirical distributions according to the topic proportions and corresponding topic embeddings. To fit such hierarchies, we develop an upward-downward optimizing strategy under the recent conditional transport theory, where document information is first transported via the upward path, and then its hierarchical representations are refined by the downward path under the Bayesian perspective. Extensive experiments on text corpora show that our approach enjoys superior modeling accuracy and interpretability. Moreover, we also conduct experiments on learning hierarchical visual topics from images, which demonstrate the adaptability and flexibility of our method.

****

On the Importance of Calibration in Semi-supervised Learning

Charlotte Loh,Rumen Dangovski,Shivchander Sudalairaj,Seungwook Han,Ligong Han,Leonid Karlinsky,Marin Soljacic,Akash Srivastava

State-of-the-art (SOTA) semi-supervised learning (SSL) methods have been highly successful in leveraging a mix of labeled and unlabeled data by combining techniques of consistency regularization and pseudo-labeling. During pseudo-labeling, the model's predictions on unlabeled data are used for training and thus, model calibration is important in mitigating confirmation bias. Yet, many SOTA methods are optimized for model performance, with little focus directed to improve model calibration. In this work, we empirically demonstrate that model calibration is strongly correlated with model performance and propose to improve calibration via approximate Bayesian techniques.

We introduce a family of new SSL models that optimizes for calibration and demonstrate their effectiveness across standard vision benchmarks of CIFAR-10, CIFAR-100 and ImageNet, giving up to 15.9\% improvement in test accuracy. Furthermore, we also demonstrate their effectiveness in additional realistic and challenging problems, such as class-imbalanced datasets and in photonics science.

****

SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-supervised Learning

Hao Chen,Ran Tao,Yue Fan,Yidong Wang,Jindong Wang,Bernt Schiele,Xing Xie,Bhiksha Raj,Marios Savvides

The critical challenge of Semi-Supervised Learning (SSL) is how to effectively l everage the limited labeled data and massive unlabeled data to improve the model's generalization performance. In this paper, we first revisit the popular pseudo-labeling methods via a unified sample weighting formulation and demonstrate the inherent quantity-quality trade-off problem of pseudo-labeling with thresholding, which may prohibit learning. To this end, we propose SoftMatch to overcome the trade-off by maintaining both high quantity and high quality of pseudo-labels during training, effectively exploiting the unlabeled data. We derive a truncated Gaussian function to weight samples based on their confidence, which can be viewed as a soft version of the confidence threshold. We further enhance the utilization of weakly-learned classes by proposing a uniform alignment approach. In experiments, SoftMatch shows substantial improvements across a wide variety of benchmarks, including image, text, and imbalanced classification.

****

Certifiably Robust Policy Learning against Adversarial Multi-Agent Communication

Yanchao Sun,Ruijie Zheng,Parisa Hassanzadeh,Yongyuan Liang,Soheil Feizi,Sumitra Ganesh,Furong Huang

Communication is important in many multi-agent reinforcement learning (MARL) problems for agents to share information and make good decisions. However, when deploying trained communicative agents in a real-world application where noise and potential attackers exist, the safety of communication-based policies becomes a severe issue that is underexplored. Specifically, if communication messages are manipulated by malicious attackers, agents relying on untrustworthy communication may take unsafe actions that lead to catastrophic consequences. Therefore, it is crucial to ensure that agents will not be misled by corrupted communication, while still benefiting from benign communication. In this work, we consider an environment with N agents, where the attacker may arbitrarily change the communication from any $C<\frac{N-1}{2}$ agents to a victim agent. For this strong threat model, we propose a certifiable defense by constructing a message-ensemble policy that aggregates multiple randomly ablated message sets. Theoretical analysis shows that this message-ensemble policy can utilize benign communication while being certifiably robust to adversarial communication, regardless of the attacking algorithm. Experiments in multiple environments verify that our defense significantly improves the robustness of trained policies against various types of attacks.

****

Node Importance Specific Meta Learning in Graph Neural Networks

Hao Liu,Yixin Chen,Dacheng Tao,Muhan Zhang

While current node classification methods for graphs have enabled significant progress in many applications, they rely on abundant labeled nodes for training. In many real-world datasets, nodes for some classes are always scarce, thus current algorithms are ill-equipped to handle these few-shot node classes. Some meta learning approaches for graphs have demonstrated advantages in tackling such few-shot problems, but they disregard the impact of node importance on a task. Being exclusive to graph data, the dependencies between nodes convey vital information for determining the importance of nodes in contrast to node features only, which poses unique challenges here. In this paper, we investigate the effect of node importance in node classification meta learning tasks. We first theoretically analyze the influence of distinguishing node importance on the lower bound of the model accuracy. Then, based on the theoretical conclusion, we propose a novel Node Importance Meta Learning architecture (NIML) that learns and applies the importance score of each node for meta learning. Specifically, after constructing an attention vector based on the interaction between a node and its neighbors, we train an importance predictor in a supervised manner to capture the distance between node embedding and the expectation of same-class embedding. Extensive experiments on public datasets demonstrate the state-of-the-art performance of NIML on few-shot node classification problems.

****

Attention-Guided Backdoor Attacks against Transformers

Weimin Lyu,Songzhu Zheng,Haibin Ling,Chao Chen

With the popularity of transformers in natural language processing (NLP) applica tions, there are growing concerns about their security. Most existing NLP attack methods focus on injecting stealthy trigger words/phrases. In this paper, we focus on the interior structure of neural networks and the Trojan mechanism. Focusing on the prominent NLP transformer models, we propose a novel Trojan Attention Loss (TAL), which enhances the Trojan behavior by directly manipulating the attention pattern. Our loss significantly improves the attack efficacy; it achieves better successful rates and with a much smaller poisoning rate (i.e., a smaller proportion of poisoned samples). It boosts attack efficacy for not only traditional dirty-label attacks, but also the more challenging clean-label attacks. TAL is also highly compatible with most existing attack methods and its flexibility enables this loss easily adapted to other backbone transformer models.

****

Disentangling the Mechanisms Behind Implicit Regularization in SGD

Zachary Novack,Simran Kaur,Tanya Marwah,Saurabh Garg,Zachary Chase Lipton

A number of competing hypotheses have been proposed to explain why small-batch Stochastic Gradient Descent (SGD) leads to improved generalization over the full-batch regime, with recent work crediting the implicit regularization of various quantities throughout training. However, to date, empirical evidence assessing the explanatory power of these hypotheses is lacking. In this paper, we conduct an extensive empirical evaluation, focusing on the ability of various theorized mechanisms to close the small-to-large batch generalization gap. Additionally, we characterize how the quantities that SGD has been claimed to (implicitly) regularize change over the course of training. By using micro-batches, i.e. disjoint smaller subsets of each mini-batch, we empirically show that explicitly penalizing the gradient norm or the Fisher Information Matrix trace, averaged over micro-batches, in the large-batch regime recovers small-batch SGD generalization, whereas Jacobian-based regularizations fail to do so. This generalization performance is shown to often be correlated with how well the regularized model’s gradient norms resemble those of small-batch SGD. We additionally show that this behavior breaks down as the micro-batch size approaches the batch size. Finally, we note that in this line of inquiry, positive experimental findings on CIFAR10 are often reversed on other datasets like CIFAR100, highlighting the need to test hypotheses on a wider collection of datasets.

****

Oracles and Followers: Stackelberg Equilibria in Deep Multi-Agent Reinforcement Learning

Matthias Gerstgrasser,David C. Parkes

Stackelberg equilibria arise naturally in a range of popular learning problems, such as in security games or indirect mechanism design, and have received in- creasing attention in the reinforcement learning literature. We present a general framework for implementing Stackelberg equilibria search as a multi-agent RL problem, allowing a wide range of algorithmic design choices. We discuss how previous approaches can be seen as specific instantiations of this framework. As a key insight, we note that the design space allows for approaches not previously seen in the literature, for instance by leveraging multitask and meta-RL techniques for follower convergence. We propose one such approach using contextual poli- cies and evaluate it experimentally on standard benchmark domains. Finally, we illustrate the effect of adopting designs outside the borders of our framework in controlled experiments.

****

Structural Code Representation Learning for Auto-Vectorization

Yao Xiao,Nesreen Ahmed,Mihai Capot■,Guixiang Ma,Theodore L. Willke,Shahin Nazarian,Paul Bogdan

The single instruction multiple data (SIMD) capability in modern processors is critical to improving the performance of current compute-intensive programs. SIMD allows architectures to exploit the natural data parallelism that exists in a wide-range of real applications (e.g., games, signal processing, etc) by executing a single instruction on multiple data items simultaneously. Modern compilers use vectorization techniques to exploit the SIMD capability, by detecting data parallelism in scalar source code and transforming a group of scalar instructions into vector-based instructions. In this work, we focus on one of the most common vectorization techniques called \emph{loop-based vectorization}, which targets loops and optimize their performance by grouping multiple occurrences of the same operation across loop iterations into single SIMD instructions. This is achieved by setting two key parameters: (1) the vectorization factor (VF), and (2) the interleaving factor (IF). Unfortunately, vectorizing loop computations effectively is a key challenging problem for both programmers and compilers due to the large search space. For example, manual vectorization of each loop puts a huge burden on the programmer, is more error-prone, and/or requires expert knowledge of both the software and the architecture. Alternatively, current compilers use fixed-cost models based on expert heuristics to make automatic vectorization decisions. However, these models often ignore the data dependencies, as well as the underlying computation graph. In this paper, we propose a data-driven graph-based learning framework for automatic vectorization, called \emph{autograph}, which takes an input program, extracts the loops, then learns a structured representation to automatically predict the correct VF/IF factors. Our proposed framework utilizes deep reinforcement learning to learn an optimal policy (observations to actions) from an intelligent agent in a SIMD environment, and automatically injects the predicted vectorization pragmas into the input program. We conducted an extensive evaluation on multiple benchmark datasets and comparisons with state-of-the-art baselines. Our experimental results show that the proposed framework can achieve up to 1.02x-2.26x and 1.06x-4.27x performance improvement, compared to state-of-the-art baseline and LLVM -O3 respectively.

****

Overthinking the Truth: Understanding how Language Models process False Demonstrations

Danny Halawi,Jean-Stanislas Denain,Jacob Steinhardt

Through few-shot learning or chain-of-thought prompting, modern language models can detect and imitate complex patterns in their prompt.
This behavior allows language models to complete challenging tasks without fine-tuning,
but can be at odds with completion quality: if the context is inaccurate or harmful, then the model
may reproduce these defects in its completions.
In this work, we show that this {harmful context-following} appears late in a model's
computation--in particular, given an inaccurate context, models perform \emph{better} after zeroing out later layers.
More concretely, at early layers models have similar performance given either accurate and inaccurate few-shot prompts, but a gap appears at later layers (e.g.~layers 10-14 for GPT-J).
This gap appears at a consistent depth across datasets, and coincides with the appearance of “induction heads” that attend to previous answers in the prompt.
We restore the performance for inaccurate contexts by ablating a subset of these heads, reducing the gap by 28\% on average across 8 datasets.
Our results suggest that studying early stages of computation could be a promising strategy to prevent misleading outputs, and that understanding and editing internal
mechanisms can help correct unwanted model behavior.

****

Sequential Attention for Feature Selection

Taisuke Yasuda,Mohammadhossein Bateni,Lin Chen,Matthew Fahrbach,Gang Fu,Vahab Mirrokni

Feature selection is the problem of selecting a subset of features for a machine learning model that maximizes model quality subject to a budget constraint. For neural networks, prior methods, including those based on ℓ_1 regularization, attention, and other techniques, typically select the entire feature subset in one evaluation round, ignoring the residual value of features during selection, i.e., the marginal contribution of a feature given that other features have already been selected. We propose a feature selection algorithm called Sequential

Attention that achieves state-of-the-art empirical results for neural networks. This algorithm is based on an efficient one-pass implementation of greedy forward selection and uses attention weights at each step as a proxy for feature importance. We give theoretical insights into our algorithm for linear regression by showing that an adaptation to this setting is equivalent to the classical Orthogonal Matching Pursuit (OMP) algorithm, and thus inherits all of its provable guarantees. Our theoretical and empirical analyses offer new explanations towards the effectiveness of attention and its connections to overparameterization, which may be of independent interest.

****

Trusted Aggregation (TAG): Model Filtering Backdoor Defense In Federated Learning

Joseph Lavond,Minhao Cheng,Yao Li

Federated Learning is a framework for training machine learning models from multiple local data sets without access to the data in aggregate. A shared model is jointly learned through an interactive process between server and clients that combines locally learned model gradients or weights. However, the lack of data transparency naturally raises concerns about model security. Recently, several state-of-the-art backdoor attacks have been proposed, which achieve high attack success rates while simultaneously being difficult to detect, leading to compromised federated learning models. In this paper, motivated by differences in the output layer distribution between models trained with and without the presence of backdoor attacks, we propose a defense method that can prevent backdoor attacks from influencing the model while maintaining the accuracy of the original classification task. TAG leverages a small validation data set to estimate the largest change that a benign user's local training can make to the output layer of the shared model, which can be used as a cutoff for returning user models. Experimental results on multiple data sets show that TAG defends against backdoor attacks even when 40\% of the user submissions to update the shared model are malicious.

****

What Deep Representations Should We Learn? -- A Neural Collapse Perspective

Xiao Li,Sheng Liu,Jinxin Zhou,Carlos Fernandez-Granda,Zhihui Zhu,Qing Qu

For classification problems, when sufficiently large networks are trained until convergence, an intriguing phenomenon has recently been discovered in the last-layer classifiers, and features termed neural collapse (NC): (i) the intra-class variability of the features collapses to zero, and (ii) the between-class feature means are maximally and equally separated. Despite of recent endeavors to understand why NC happens, a fundamental question remains: whether NC is a blessing or a curse for deep learning? In this work, we investigate the problem under the setting of transfer learning that we pretrain a model on a large dataset and transfer it to downstream tasks. Through various experiments, our findings on NC are two-fold: (i) when pretrain models, preventing intra-class variability collapse (to a certain extent) better preserve the structures of data, and leads to better model transferability; (ii) when fine-tuning models on downstream tasks, obtaining features with more NC on downstream data results in better test accuracy on the given task. Our findings based upon NC not only explain many widely used heuristics in model pretraining (e.g., data augmentation, projection head, self-supervised learning), but also leads to more efficient and principled transfer learning method on downstream tasks.

****

Improved Sample Complexity for Reward-free Reinforcement Learning under Low-rank MDPs

Yuan Cheng,Ruiquan Huang,Yingbin Liang,Jing Yang

In reward-free reinforcement learning (RL), an agent explores the environment first without any reward information, in order to achieve certain learning goals afterwards for any given reward. In this paper we focus on reward-free RL under low-rank MDP models, in which both the representation and linear weight vectors are unknown. Although various algorithms have been proposed for reward-free low-rank MDPs, the corresponding sample complexity is still far from being satisfactory. In this work, we first provide the first known sample complexity lower bound that holds for any algorithm under low-rank MDPs. This lower bound implies it is strictly harder to find a near-optimal policy under low-rank MDPs than under linear MDPs. We then propose a novel model-based algorithm, coined RAFFLE, and show it can both find an ϵ-optimal policy and achieve an ϵ-accurate system identification via reward-free exploration, with a sample complexity significantly improving the previous results. Such a sample complexity matches our lower bound in the dependence on ϵ, as well as on K {in the large d regime}, where d and K respectively denote the representation dimension and action space cardinality. Finally, we provide a planning algorithm (without further interaction with true environment) for RAFFLE to learn a near-accurate representation, which is the first known representation learning guarantee under the same setting.

****

Re-Imagen: Retrieval-Augmented Text-to-Image Generator
Wenhu Chen,Hexiang Hu,Chitwan Saharia,William W. Cohen

Research on text-to-image generation has witnessed significant progress in generating diverse and photo-realistic images, driven by diffusion and auto-regressive models trained on large-scale image-text data. Though state-of-the-art models can generate high-quality images of common entities, they often have difficulty generating images of uncommon entities, such as `Chortai (dog)' or `Picarones (food)'. To tackle this issue, we present the Retrieval-Augmented Text-to-Image Generator (Re-Imagen), a generative model that uses retrieved information to produce high-fidelity and faithful images, even for rare or unseen entities. Given a text prompt, Re-Imagen accesses an external multi-modal knowledge base to retrieve relevant (image, text) pairs, and uses them as references to generate the image. With this retrieval step, Re-Imagen is augmented with the knowledge of high-level semantics and low-level visual details of the mentioned entities, and thus improves its accuracy in generating the entities' visual appearances. We train Re-Imagen on a constructed dataset containing (image,text,retrieval) triples to teach the model to ground on both text prompt and retrieval. Furthermore, we develop a new sampling strategy to interleave the classifier-free guidance for text and retrieval condition to balance the text and retrieval alignment. Re-Imagen achieves new SoTA FID results on two image generation benchmarks, such as COCO (\ie, FID = 5.25) and WikiImage (\ie, FID = 5.82) without fine-tuning. To further evaluate the capabilities of the model, we introduce EntityDrawBench, a new benchmark that evaluates image generation for diverse entities, from frequent to rare, across multiple visual domains. Human evaluation on EntityDrawBench shows that Re-Imagen performs on par with the best prior models in photo-realism, but with significantly better real-world faithfulness, especially on less frequent entities.

****

Provably Efficient Lifelong Reinforcement Learning with Linear Representation
Sanae Amani,Lin Yang,Ching-An Cheng

We theoretically study lifelong reinforcement learning (RL) with linear representation in a regret minimization setting. The goal of the agent is to learn a multi-task policy based on a linear representation while solving a sequence of tasks that may be adaptively chosen based on the agent's past behaviors. We frame the problem as a linearly parameterized contextual Markov decision process (MDP), where each task is specified by a context and the transition dynamics is context-independent, and we introduce a new completeness-style assumption on the representation which is sufficient to ensure the optimal multi-task policy is realizable under the linear representation. Under this assumption, we propose an algorithm, called UCB Lifelong Value Distillation (UCBlvd), that provably achieves sublinear regret for any sequence of tasks while using only sublinear planning calls. Specifically, for K task episodes of horizon H, our algorithm has a regret bound $\tilde{\mathcal{O}}(\sqrt{(d^3+d^\prime d)H^4K})$ using $\mathcal{O}(dH\log(K))$ number of planning calls, where d and $d^\prime$ are the feature dimensions of the dynamics and rewards, respectively. This theoretical guarantee implies that our algorithm can enable a lifelong learning agent to learn to interna lize experiences into a multi-task policy and rapidly solve new tasks.

****

Towards Adversarially Robust Deepfake Detection: An Ensemble Approach
Ashish Hooda,Neal Mangaokar,Ryan Feng,Kassem Fawaz,Somesh Jha,Atul Prakash
Detecting deepfakes remains an open problem. Current detection methods fail against an adversary who adds imperceptible adversarial perturbations to the deepfake to evade detection. We propose Disjoint Deepfake Detection (D3), a deepfake detector designed to improve adversarial robustness beyond de facto solutions such as adversarial training. D3 uses an ensemble of models over disjoint subsets of the frequency spectrum to significantly improve robustness. Our key insight is to leverage a redundancy in the frequency domain and apply a saliency partitioning technique to disjointly distribute frequency components across multiple models. We formally prove that these disjoint ensembles lead to a reduction in the dimensionality of the input subspace where adversarial deepfakes lie. We then empirically validate the D3 method against white-box attacks and black-box attacks and find that D3 significantly outperforms existing state-of-the-art defenses applied to deepfake detection.

****

Fast Adaptation via Human Diagnosis of Task Distribution Shift
Andi Peng,Mark K Ho,Aviv Netanyahu,Julie Shah,Pulkit Agrawal
When agents fail in the world, it is important to understand why they failed. These errors could be due to underlying distribution shifts in the goals desired by the end user or to the environment layouts that impact the policy's actions. In the case of multi-task policies conditioned on goals, this problem manifests in difficulty in disambiguating between goal and policy failures: is the agent failing because it can't correctly infer what the desired goal is or because it doesn't know how to take actions toward achieving the goal? We hypothesize that successfully disentangling these two failures modes holds important implications for selecting a finetuning strategy. In this paper, we explore the feasibility of leveraging human feedback to diagnose what vs. how failures for efficient adaptation. We develop an end-to-end policy training framework that uses attention to produce a human-interpretable representation, a visual masked state, to communicate the agent's intermediate task representation. In experiments with human users in both discrete and continuous control domains, we show that our visual attention mask policy can aid participants in successfully inferring the agent's failure mode significantly better than actions alone. Leveraging this feedback, we show subsequent empirical performance gains during finetuning and discuss implications of using humans to diagnose parameter-level failures of distribution shift.

****

Link Prediction with Non-Contrastive Learning
William Shiao,Zhichun Guo,Tong Zhao,Evangelos E. Papalexakis,Yozen Liu,Neil Shah
Graph neural networks (GNNs) are prominent in the graph machine learning domain, owing to their strong performance across various tasks. A recent focal area is the space of graph self-supervised learning (SSL), which aims to derive useful node representations without labeled data. Notably, many state-of-the-art graph SSL methods are contrastive methods, which use a combination of positive and negative samples to learn node representations. Owing to challenges in negative sampling (slowness and model sensitivity), recent literature introduced non-contrastive methods, which instead only use positive samples. Though such methods have shown promising performance in node-level tasks, their suitability for link prediction tasks, which are concerned with predicting link existence between pairs of nodes (and have broad applicability to recommendation systems contexts) is yet unexplored. In this work, we extensively evaluate the performance of existing non-contrastive methods for link prediction in both transductive and inductive settings. While most existing non-contrastive methods perform poorly overall, we find that, surprisingly, BGRL generally performs well in transductive settings. However, it performs poorly in the more realistic inductive settings where the model has to generalize to links to/from unseen nodes. We find that non-contrastive models tend to overfit to the training graph and use this analysis to propose T

-BGRL, a novel non-contrastive framework that incorporates cheap corruptions to improve the generalization ability of the model. This simple modification strongly improves inductive performance in 5/6 of our datasets, with up to a 120% improvement in Hits@50 - all with comparable speed to other non-contrastive baselines, and up to $14\times$ faster than the best-performing contrastive baseline. Our work imparts interesting findings about non-contrastive learning for link prediction and paves the way for future researchers to further expand upon this area.

****

Distributed Differential Privacy in Multi-Armed Bandits

Sayak Ray Chowdhury,Xingyu Zhou

We consider the standard K-armed bandit problem under a distributed trust model of differential privacy (DP), which enables to guarantee privacy without a trustworthy server. Under this trust model, previous work largely focus on achieving privacy using a shuffle protocol, where a batch of users data are randomly permuted before sending to a central server. This protocol achieves (ϵ,δ) or approximate-DP guarantee by sacrificing an additive $O\!\left(\!\frac{K\log T\sqrt{\log(1/\delta)}}{\epsilon}\!\right)\!$ factor in T-step cumulative regret. In contrast, the optimal privacy cost to achieve a stronger ($\epsilon,0$) or pure-DP guarantee under the widely used central trust model is only $\Theta\!\left(\!\frac{K\log T}{\epsilon}\!\right)\!$, where, however, a trusted server is required. In this work, we aim to obtain a pure-DP guarantee under distributed trust model while sacrificing no more regret than that under central trust model. We achieve this by designing a generic bandit algorithm based on successive arm elimination, where privacy is guaranteed by corrupting rewards with an equivalent discrete Laplace noise ensured by a secure computation protocol. We also show that our algorithm, when instantiated with Skellam noise and the secure protocol, ensures \emph{R\'{e}nyi differential privacy} -- a stronger notion than approximate DP -- under distributed trust model with a privacy cost of $O\!\left(\!\frac{K\sqrt{\log T}}{\epsilon}\!\right)\!$. Finally, as a by-product of our techniques, we also recover the best-known regret bounds for bandits under central and local models while using only \emph{discrete privacy noise}, which can avoid the privacy leakage due to floating point arithmetic of continuous noise on finite computers.

****

Thrust: Adaptively Propels Large Language Models with External Knowledge

Xinran Zhao,Hongming Zhang,Xiaoman Pan,Wenlin Yao,Dong Yu,Jianshu Chen

Large-scale pre-trained language models (PTLM) have achieved great success in various natural language processing (NLP) tasks. Much evidence shows that PTLMs already encode rich knowledge themselves, but knowledge stored in PTLMs can be opaque and static, making external knowledge retrieval necessary. However, there are two major challenges when using external knowledge. First, knowledge indexing and retrieving on large-scale knowledge bases are time costly. Second, knowledge retrieved could be noisy and sometimes misleading. Motivated by the observation that external knowledge is not always required by PTLMs, we investigate an effective and efficient way to apply knowledge only when the knowledge is essential. Specifically, we propose instance-level adaptive propulsion of external knowledge (IAPEK), where we score each instance on whether the PTLMs need the support of external knowledge. To achieve this goal, we design a novel metric, Thrust, which leverages the distribution estimation on seen/training instances. Extensive experiments demonstrate that we can achieve significantly higher cost-efficiency through Thrust compared to the naive usage of external knowledge on 88% of the evaluated tasks with 26% average performance improvement. Such findings further shed light on the real-world practice of knowledge-enhanced LMs with a limited budget for knowledge seeking due to computation latency or costs.

****

Progress measures for grokking via mechanistic interpretability

Neel Nanda,Lawrence Chan,Tom Lieberum,Jess Smith,Jacob Steinhardt

Neural networks often exhibit emergent behavior in which qualitatively new capabilities that arise from scaling up the number of parameters, training data, or e ven the number of steps. One approach to understanding emergence is to find the continuous \textit{progress measures} that underlie the seemingly discontinuous qualitative changes. In this work, we argue that progress measures can be found via mechanistic interpretability---that is, by reverse engineering learned models into components and measuring the progress of each component over the course of training. As a case study, we study small transformers trained on a modular arithmetic tasks with emergent grokking behavior. We fully reverse engineer the algorithm learned by these networks, which uses discrete fourier transforms and trigonometric identities to convert addition to rotation about a circle. After confirming the algorithm via ablation, we then use our understanding of the algorithm to define progress measures that precede the grokking phase transition on this task. We see our result as demonstrating both that it is possible to fully reverse engineer trained networks, and that doing so can be invaluable to understanding their training dynamics.

****

On the Mysterious Optimization Geometry of Deep Neural Networks

Chedi Morchdi,Yi Zhou,Jie Ding,Bei Wang

Understanding why gradient-based algorithms are successful in practical deep learning optimization is a fundamental and long-standing problem. Most existing works promote the explanation that deep neural networks have smooth and amenable nonconvex optimization geometries. In this work, we argue that this may be an oversimplification of practical deep learning optimization by revealing a mysterious and complex optimization geometry of deep networks through extensive experiments. Specifically, we consistently observe two distinct geometric patterns in training various deep networks: a regular smooth geometry and a mysterious zigzag geometry, where gradients computed in adjacent iterations are extremely negatively correlated. Also, such a zigzag geometry exhibits a fractal structure in that it appears over a wide range of geometrical scales, implying that deep networks can be highly non-smooth in certain local parameter regions. Moreover, our results show that a substantial part of the training progress is achieved under such complex geometry. Therefore, the existing smoothness-based explanations do not fully match the practice.

****

Implicit regularization via Spectral Neural Networks and non-linear matrix sensing

Subhroshekhar Ghosh,Thanh Lam,Soumendu Sundar Mukherjee

The phenomenon of \textit{implicit regularization} has attracted interest in the recent years as a fundamental aspect of the remarkable generalizing ability of neural networks. In a nutshell, it entails that gradient flow dynamics in many neural nets, even without any explicit regularizer in the loss function, converges to the solution of a regularized learning problem. However, known results attempting to theoretically explain this phenomenon focus overwhelmingly on the setting of linear neural nets, and the simplicity of the linear structure is particularly crucial to existing arguments. In this paper, we explore this problem in the context of more realistic neural networks with a general class of non-linear activation functions, and rigorously demonstrate the implicit regularization phenomenon for such networks in the setting of matrix sensing problems. This is coupled with rigorous rate guarantees that ensure exponentially fast convergence of gradient descent, complemented by matching lower bounds which stipulate that the exponential rate is the best achievable. In this vein, we contribute a network architecture called Spectral Neural Networks (\textit{abbrv.} SNN) that is particularly suitable for matrix learning problems. Conceptually, this entails coordinatizing the space of matrices by their singular values and singular vectors, as opposed to by their entries, a potentially fruitful perspective for matrix learning. We demonstrate that the SNN architecture is inherently much more amenable to theoretical analysis than vanilla neural nets and confirm its effectiveness in the context of matrix sensing, supported via both mathematical guarantees and empirical investigations. We believe that the SNN architecture has the potential to be of wide applicability in a broad class of matrix learning scenarios.

****

Goal-Space Planning with Subgoal Models
Chunlok Lo,Gabor Mihucz,Farzane Aminmansour,Adam White,Martha White

This paper investigates a new approach to model-based reinforcement learning using background planning: mixing (approximate) dynamic programming updates and model-free updates, similar to the Dyna architecture. Background planning with learned models is often worse than model-free alternatives, such as Double DQN, even though the former uses significantly more memory and computation. The fundamental problem is that learned models can be inaccurate and often generate invalid states, especially when iterated many steps. In this paper, we avoid this limitation by constraining background planning to a set of (abstract) subgoals and learning only local, subgoal-conditioned models. This goal-space planning (GSP) approach is more computationally efficient, naturally incorporates temporal abstraction for faster long-horizon planning and avoids learning the transition dynamics entirely. We show that our GSP algorithm can learn significantly faster than a Double DQN baseline in a variety of situations.

****

MET : Masked Encoding for Tabular data
Kushal Alpesh Majmundar,Sachin Goyal,Praneeth Netrapalli,Prateek Jain

We propose $\textit{Masked Encoding for Tabular Data (MET)}$ for learning self-supervised representations from $\textit{tabular data}$. Tabular self-supervised learning (tabular-SSL) -- unlike structured domains like images, audio, text -- is more challenging since each tabular dataset can have a completely different structure among its features (or coordinates), which is hard to identify a priori. $\textit{MET}$ attempts to circumvent this problem by assuming the following hypothesis: the observed tabular data features come from a latent graphical model and the downstream tasks are significantly easier to solve in the latent space. Based on this hypothesis, $\textit{MET}$ uses random masking based encoders to learn a positional embedding for each coordinate, which would in turn capture the latent structure between coordinates. Through experiments on a toy dataset from a linear graphical model, we show that $\textit{MET}$ is indeed able to capture the latent graphical model. Practically, through extensive experiments on multiple benchmarks for tabular data, we demonstrate that $\textit{MET}$ significantly outperforms all the baselines. For example, on Criteo -- a large-scale click prediction dataset -- $\textit{MET}$ achieves as much as 5% improvement over the current state-of-the-art (SOTA) while purely supervised learning based approaches have been able to advance SOTA by at most 2% in the last six years. Furthermore, averaged over $\textit{nine}$ datasets, $\textit{MET}$ is around 3.9% more accurate than the next best method of Gradient-boosted decision trees -- considered as SOTA for the tabular setting.

****

Shortcut Learning Through the Lens of Early Training Dynamics
Nihal Murali,Aahlad Manas Puli,Ke Yu,Rajesh Ranganath,kayhan Batmanghelich

Deep Neural Networks (DNNs) are prone to learn shortcut patterns that damage the generalization of the DNN during deployment. Shortcut learning is concerning, particularly when the DNNs are applied to safety-critical domains. This paper aims to better understand shortcut learning through the lens of the learning dynamics of the internal neurons during the training process. More specifically, we make the following observations: (1) While previous works treat shortcuts as synonymous with spurious correlations, we emphasize that not all spurious correlations are shortcuts. We show that shortcuts are only those spurious features that are ``easier'' than the core features. (2) We build upon this premise and use instance difficulty methods (like Prediction Depth) to quantify ``easy'' and to identify this behavior during the training phase. (3) We empirically show that shortcut learning can be detected by observing the learning dynamics of the DNN's early layers, irrespective of the network architecture used. In other words, easy features learned by the initial layers of a DNN early during the training are potential shortcuts. We verify our claims on simulated and real medical imaging data and justify the empirical success of our hypothesis by showing the theoretical connections between Prediction Depth and information-theoretic concepts like $\mathcal{V}$-usable information. Lastly, our experiments show the insufficiency o f monitoring only accuracy plots during training (as is common in machine learning pipelines), and we highlight the need for monitoring early training dynamics using example difficulty metrics.

****

On $\mathcal{O}(1/K)$ Convergence and Low Sample Complexity for Single-Timescale Policy Evaluation with Nonlinear Function Approximation

Zhuqing Liu,Xin Zhang,Jia Liu,Zhengyuan Zhu,Songtao Lu

■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 {\em non-linear} function approximation. This challenge motivates us to develop a new variance-reduced primal-dual method (VRPD) that is able to achieve a fast convergence speed for RL policy evaluation with nonlinear function approximation. To lower the high sample complexity limitation of variance-reduced approaches (due to the periodic full gradient evaluation with all training data), we further propose an enhanced VRPD method with an adaptive-batch adjustment (VRPD$^+$). The main features of VRPD include: i) VRPD allows the use of {\em{constant}} step sizes and achieves the $\mathcal{O}(1/K)$ convergence rate to the first-order stationary points of non-convex policy evaluation problems; ii) VRPD 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, VRPD$^+$ is more sample-efficient in practice without loss of theoretical convergence rate. Our extensive numerical experiments verify our theoretical findings and showcase the high efficiency of the proposed VRPD and VRPD$^+$ algorithms compared with the state-of-the-art methods.

****

On the Implicit Bias Towards Depth Minimization in Deep Neural Networks

Tomer Galanti,Liane Galanti,Ido Ben-Shaul

Recent results in the literature suggest that the penultimate (second-to-last) layer representations of neural networks that are trained for classification exhibit a clustering property called neural collapse (NC). We study the implicit bias of stochastic gradient descent (SGD) in favor of low-depth solutions when training deep neural networks. We characterize a notion of effective depth that measures the first layer for which sample embeddings are separable using the nearest-class center classifier. Furthermore, we hypothesize and empirically show that SGD implicitly selects neural networks of small effective depths.

Secondly, while neural collapse emerges even when generalization should be impossible - we argue that the \emph{degree of separability} in the intermediate layers is related to generalization. We derive a generalization bound based on comparing the effective depth of the network with the minimal depth required to fit the same dataset with partially corrupted labels. Remarkably, this bound provides non-trivial estimations of the test performance. Finally, we empirically show that the effective depth of a trained neural network monotonically increases when increasing the number of random labels in data.

****

Prometheus: Endowing Low Sample and Communication Complexities to Constrained Decentralized Stochastic Bilevel Learning

Zhuqing Liu,Xin Zhang,Prashant Khanduri,Songtao Lu,Jia Liu

■In recent years, constrained decentralized stochastic bilevel optimization has become increasingly important due to its versatility in modeling a wide range of multi-agent learning problems, such as multi-agent reinforcement learning and multi-agent meta-learning with safety constraints. However, one under-explored and fundamental challenge in constrained decentralized stochastic bilevel optimization is how to achieve low sample and communication complexities, which, if not addressed appropriately, could affect the long-term prospect of many emerging multi-agent learning paradigms that use decentralized bilevel optimization as a bedrock. In this paper, we investigate a class of constrained decentralized bileve l optimization problems, where multiple agents collectively solve a nonconvex-strongly-convex bilevel problem with constraints in the upper-level variables. Such problems arise naturally in many multi-agent reinforcement learning and meta learning problems. In this paper, we propose an algorithm called Prometheus (proximal tracked stochastic recursive estimator) that achieves the first $\mathcal{O}(\epsilon^{-1})$ results in both sample and communication complexities for constrained decentralized bilevel optimization, where $\epsilon>0$ is the desired stationarity error. Collectively, the results in this work contribute to a theoretical foundation for low sample- and communication-complexity constrained decentralized bilevel learning.

****

SGD and Weight Decay Provably Induce a Low-Rank Bias in Neural Networks

Tomer Galanti,Zachary S Siegel,Aparna Gupte,Tomaso Poggio

We analyze deep ReLU neural networks trained with mini-batch Stochastic Gradient Descent (SGD) and weight decay. We show, both theoretically and empirically, that when training a neural network using SGD with weight decay and small batch size, the resulting weight matrices tend to be of small rank. Our analysis relies on a minimal set of assumptions; the neural networks may be arbitrarily wide or deep and may include residual connections, as well as convolutional layers.

The same analysis implies the inherent presence of SGD ``noise'', defined as the inability of SGD to converge to a stationary point. In particular, we prove that SGD noise must always be present, even asymptotically, as long as we incorporate weight decay and the batch size is smaller than the total number of training samples.

****

MobileViTv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features

Shakti Nagnath Wadekar,Abhishek Chaurasia

MobileViT (MobileViTv1) combines convolutional neural networks (CNNs) and vision transformers (ViTs) to create light-weight models for mobile vision tasks. Though the main MobileViTv1-block helps to achieve competitive state-of-the-art results, the fusion block inside MobileViTv1-block, creates scaling challenges and has a complex learning task. We propose changes to the fusion block that are simple and effective to create MobileViTv3-block, which addresses the scaling and simplifies the learning task. Our proposed MobileViTv3-block used to create MobileViTv3-XXS, XS and S models outperform MobileViTv1 on ImageNet-1k, ADE20K, COCO and PascalVOC2012 datasets. On ImageNet-1K, MobileViTv3-XXS and MobileViTv3-XS surpasses MobileViTv1-XXS and MobileViTv1-XS by 2% and 1.9% respectively. Recently published MobileViTv2 architecture removes fusion block and uses linear complexity transformers to perform better than MobileViTv1. We add our proposed fusion block to MobileViTv2 to create MobileViTv3-0.5, 0.75 and 1.0 models. MobileViTv3-0.5 and MobileViTv3-0.75 outperforms MobileViTv2-0.5 and MobileViTv2-0.75 by 2.1% and 1.0% respectively on ImageNet-1K dataset. For segmentation task, MobileViTv3-1.0 achieves 2.07% and 1.1% better mIOU compared to MobileViTv2-1.0 on ADE20K dataset and PascalVOC2012 dataset respectively.

****

PiFold: Toward effective and efficient protein inverse folding

Zhangyang Gao,Cheng Tan,Stan Z. Li

How can we design protein sequences folding into the desired structures effectively and efficiently? Structure-based protein design has attracted increasing attention in recent years; however, few methods can simultaneously improve the accuracy and efficiency due to the lack of expressive features and autoregressive sequence decoder. To address these issues, we propose PiFold, which contains a novel residue featurizer and PiGNN layers to generate protein sequences in a one-shot way with improved recovery. Experiments show that PiFold could achieve 51.66\% recovery on CATH 4.2, while the inference speed is 70 times faster than the autoregressive competitors. In addition, PiFold achieves 58.72\% and 60.42\% recovery scores on TS50 and TS500, respectively. We conduct comprehensive ablation studies to reveal the role of different types of protein features and model designs, inspiring further simplification and improvement.

****

A Theoretical Understanding of Shallow Vision Transformers: Learning, Generalization, and Sample Complexity

Hongkang Li,Meng Wang,Sijia Liu,Pin-Yu Chen

Vision Transformers (ViTs) with self-attention modules have recently achieved great empirical success in many vision tasks. Due to non-convex interactions across layers, however, the theoretical learning and generalization analysis is mostly elusive. Based on a data model characterizing both label-relevant and label-irrelevant tokens, this paper provides the first theoretical analysis of training a three-layer ViT, i.e., one self-attention layer followed by a two-layer perceptron, for a classification task. We characterize the sample complexity to achieve a zero generalization error. Our sample complexity bound is positively correlated with the inverse of the fraction of label-relevant tokens, the token noise level, and the initial model error. We also prove that a training process using stochastic gradient descent (SGD) leads to a sparse attention map, which is a formal verification of the general intuition about the success of attention. Moreover, this paper indicates that a proper token sparsification can improve the test performance by removing label-irrelevant and/or noisy tokens, including spurious correlations. Empirical experiments on synthetic data and CIFAR-10 dataset justify our theoretical results and generalize to deeper ViTs.

****

AlphaDesign: A graph protein design method and benchmark on AlphaFold DB

Zhangyang Gao,Cheng Tan,Stan Z. Li

While AlphaFold has remarkably advanced protein folding, the inverse problem, protein design, by which protein sequences are predicted from the corresponding 3D structures, still faces significant challenges. First of all, there lacks a large-scale benchmark covering the vast protein space for evaluating methods and models; secondly, existing methods are still low in prediction accuracy and time-inefficient inference. This paper establishes a new benchmark based on AlphaFold DB, one of the world's largest protein structure databases. Moreover, we propose a new baseline method called AlphaDesign, which achieves 5\% higher recovery than previous methods and about 70 times inference speed-up in designing long protein sequences. We also reveal AlphaDesign's potential for practical protein design tasks, where the designed proteins achieve good structural compatibility with native structures. The open-source code will be released.

****

Transfer Learning with Context-aware Feature Compensation

Long Bo,Su Yang

Transfer learning aims to reuse the learnt representations or subnetworks to a new domain with minimum effort for adaption. Here, the challenge lies in the mismatch between source domain and target domain, which is the major gap to be tackled by transfer learning. Hence, how to identify the mismatch between source and target domain becomes a critical problem. We propose an end-to-end framework to learn feature compensation for transfer learning with soft gating to decide whether and how much feature compensation is needed, accounting for the mismatch between source domain and target domain. To enable identifying the position of the input in reference to the overall data distribution of source domain, we perform clustering at first to figure out the data distribution in a compact form represented by cluster centers, and then use the similarities between the input and the cluster centers to describe the relative position of the input in reference to the cluster centers. This acts as the context to indicate whether and how much feature compensation is needed for the input to compensate for the mismatch between source domain and target domain. To approach that, we add only two subnetworks in the form of Multilayer Perceptron, one for computing the feature compensation and the other for soft gating the compensation, where both are computed based on the context. The experiments show that such minor change to backbone network can result in significant performance improvements compared with the baselines on some widely used benchmarks.

****

Contrastive Learning Can Find An Optimal Basis For Approximately View-Invariant

Functions
Daniel D. Johnson,Ayoub El Hanchi,Chris J. Maddison

Contrastive learning is a powerful framework for learning self-supervised representations that generalize well to downstream supervised tasks. We show that multiple existing contrastive learning methods can be reinterpeted as learning kernel functions that approximate a fixed *positive-pair kernel*. We then prove that a simple representation obtained by combining this kernel with PCA provably minimizes the worst-case approximation error of linear predictors, under a straightforward assumption that positive pairs have similar labels. Our analysis is based on a decomposition of the target function in terms of the eigenfunctions of a positive-pair Markov chain, and a surprising equivalence between these eigenfunctions and the output of Kernel PCA. We give generalization bounds for downstream linear prediction using our kernel PCA representation, and show empirically on a set of synthetic tasks that applying kernel PCA to contrastive learning models can indeed approximately recover the Markov chain eigenfunctions, although the accuracy depends on the kernel parameterization as well as on the augmentation strength.

****

Learning Unsupervised Forward Models from Object Keypoints
Alireza Rezazadeh,Changhyun Choi

Object-centric representation is an essential abstraction for forward prediction. Most existing forward models learn this representation through extensive supervision (e.g., object class and bounding box) although such ground-truth information is not readily accessible in reality. To address this, we introduce KINet (Keypoint Interaction Network)---an end-to-end unsupervised framework to reason about object interactions based on a keypoint representation. Using visual observations, our model learns to associate objects with keypoint coordinates and discovers a graph representation of the system as a set of keypoint embeddings and their relations. It then learns an action-conditioned forward model using contrastive estimation to predict future keypoint states. By learning to perform physical reasoning in the keypoint space, our model automatically generalizes to scenarios with a different number of objects, novel backgrounds, and unseen object geometries. Experiments demonstrate the effectiveness of our model in accurately performing forward prediction and learning plannable object-centric representations which can also be used in downstream robotic manipulation tasks.

****

K-SAM: Sharpness-Aware Minimization at the Speed of SGD
Renkun Ni,Ping-yeh Chiang,Jonas Geiping,Micah Goldblum,Andrew Gordon Wilson,Tom Goldstein

Sharpness-Aware Minimization (SAM) has recently emerged as a robust technique for improving the accuracy of deep neural networks. However, SAM incurs a high computational cost in practice, requiring up to twice as much computation as vanilla SGD. The computational challenge posed by SAM arises because each iteration requires both ascent and descent steps and thus double the gradient computations. To address this challenge, we propose to compute gradients in both stages of SAM on only the top-k samples with highest loss. K-SAM is simple and extremely easy-to-implement while providing significant generalization boosts over vanilla SGD at little to no additional cost.

****

Copula Conformal Prediction for Multi-step Time Series Forecasting
Sophia Huiwen Sun,Rose Yu

Accurate uncertainty measurement is a key step to building robust and reliable machine learning systems. Conformal prediction is a distribution-free uncertainty quantification algorithm popular for its ease of implementation, statistical coverage guarantees, and versatility for underlying forecasters. However, existing conformal prediction algorithms for time series are limited to single-step prediction without considering the temporal dependency. In this paper we propose a Copula-based Conformal Prediction algorithm for multivariate, multi-step Time Series forecasting, CopulaCPTS. On several synthetic and real-world multivariate time series datasets, we show that CopulaCPTS produces more calibrated and sharp c onfidence intervals for multi-step prediction tasks than existing techniques.

****

Multi-Epoch Matrix Factorization Mechanisms for Private Machine Learning

Christopher A. Choquette-Choo,Hugh Brendan McMahan,J Keith Rush,Abhradeep Guha Thakurta

We introduce new differentially private (DP) mechanisms for gradient-based machine learning (ML) training involving multiple passes (epochs) of a dataset, substantially improving the achievable privacy-utility-computation tradeoffs. Our key contribution is an extension of the online matrix factorization DP mechanism to multiple participations, substantially generalizing the approach of DMRST2022. We first give a non-trivial reduction of the problem with per-iteration vector contributions to the simpler one of scalar contributions. Using this, we formulate the construction of optimal (in total squared error at each iterate) matrix mechanisms for SGD variants as a convex program. We provide a closed form solution to the dual function, leading directly to an efficient optimization algorithms.

While tractable, both solving the convex problem offline and computing the necessary noise masks during training can become prohibitively expensive when many training steps are necessary. To address this, we design a Fourier-transform-based mechanism with significantly less computation and only a minor utility decrease.

Extensive empirical evaluation on two tasks, example-level DP for image classification and user-level DP for language modeling, demonstrate substantial improvements over the previous state-of-the-art. Though our primary application is to ML, we note our main DP results are applicable to arbitrary linear queries and hence may have much broader applicability.

****

Quantum 3D graph structure learning with applications to molecule computing

Ge Yan,Huaijin Wu,Junchi Yan

Graph representation learning has been extensively studied over the last decades, and recent models start to focus on an under-explored area of 3D graph learning with 3D spatial position as well as node attributes. Despite the progress, the ability to understand the physical meaning of the 3D topology data is still a bottleneck for existing models. On the other hand, quantum computing is known to be a promising direction for theoretically verified supremacy as well as increasing evidence for access to a physical quantum device in the near term. For the first time, we propose a quantum 3D embedding ansatz that learns the latent representation of 3D structures from the Hilbert space composed of the Bloch sphere of each qubit. We convert the 3D Cartesian coordinates of nodes into rotation and torsion angles and then encode them into the form of qubits. Moreover, Parameterized Quantum Circuit (PQC) is applied to serve as the trainable layers and we take the output of the PQC as the node embedding. Experimental results on two downstream tasks, molecular property prediction and 3D molecular geometries generation, demonstrate the effectiveness of our model. Though the results are still restricted by computational power, we have shown the capability of our model with very few parameters and the potential to execute on a real quantum device.

****

Distributional Signals for Node Classification in Graph Neural Networks

Feng Ji,See Hian Lee,Zhao Kai,Wee Peng Tay,Jielong Yang

In graph neural networks (GNNs), both node features and labels are examples of graph signals, a key notion in graph signal processing (GSP). While it is common in GSP to impose signal smoothness constraints in learning and estimation tasks, it is unclear how this can be done for discrete node labels. We bridge this gap by introducing the concept of distributional graph signals. In our framework, we work with the distributions of node labels instead of their values and propose notions of smoothness and non-uniformity of such distributional graph signals. We then propose a general regularization method for GNNs that allows us to encode distributional smoothness and non-uniformity of the model output in semi-supervised node classification tasks. Numerical experiments demonstrate that our meth od can significantly improve the performance of most base GNN models in different problem settings.

****

Vector Quantized Wasserstein Auto-Encoder

Long Tung Vuong,Trung Le,He Zhao,Chuanxia Zheng,Mehrtash Harandi,Jianfei Cai,Dinh Phung

Learning deep discrete latent presentations offers a promise of better symbolic and summarized abstractions that are more useful to subsequent downstream tasks. Recent work on Vector Quantized Variational Auto-Encoder (VQ-VAE) has made substantial progress in this direction. However, this quantizes latent representations using the online k-means algorithm which suffers from poor initialization and non-stationary clusters. To strengthen the clustering quality for the latent representations, we propose Vector Quantized Wasserstein Auto-Encoder (VQ-WAE) intuitively developed based on the clustering viewpoint of Wasserstein (WS) distance. Specifically, we endow a discrete distribution over the codewords and learn a deterministic decoder that transports the codeword distribution to the data distribution via minimizing a WS distance between them. We develop further theories to connect it with the clustering viewpoint of WS distance, allowing us to have a better and more controllable clustering solution. Finally, we empirically evaluate our method on several well-known benchmarks, where it achieves better qualitative and quantitative performances than the baselines in terms of the codebook utilization and image reconstruction/generation.

****

Exact Representation of Sparse Networks with Symmetric Nonnegative Embeddings

Sudhanshu Chanpuriya,Ryan A. Rossi,Anup Rao,Tung Mai,Nedim Lipka,Zhao Song,Cameron N Musco

Many models for undirected graphs are based on factorizing the graph's adjacency matrix; these models find a vector representation of each node such that the predicted probability of a link between two nodes increases with the similarity (dot product) of their associated vectors. Recent work has shown that these models are unable to capture key structures in real-world graphs, particularly heterophilous structures, wherein links occur between dissimilar nodes. In contrast, a factorization with two vectors per node, based on logistic principal components analysis (LPCA), has been proven not only to represent such structures, but also to provide exact low-rank factorization of any graph with bounded max degree. However, this bound has limited applicability to real-world networks, which often have power law degree distributions with high max degree. Further, the LPCA model lacks interpretability since its asymmetric factorization does not reflect the undirectedness of the graph. We address the above issues in two ways. First, we prove a new bound for the LPCA model in terms of arboricity rather than max degree; this greatly increases the bound's applicability to many sparse real-world networks. Second, we propose an alternative graph model whose factorization is symmetric and nonnegative, which allows for link predictions to be interpreted in terms of node clusters. We show that the bounds for exact representation in the LPCA model extend to our new model. On the empirical side, our model is optimized effectively on real-world graphs with gradient descent on a cross-entropy loss. We demonstrate its effectiveness on a variety of foundational tasks, such as community detection and link prediction.

****

Skill-Based Reinforcement Learning with Intrinsic Reward Matching

Ademi Adeniji,Amber Xie,Pieter Abbeel

While unsupervised skill discovery has shown promise in autonomously acquiring behavioral primitives, there is still a large methodological disconnect between task-agnostic skill pretraining and downstream, task-aware finetuning. We present Intrinsic Reward Matching (IRM), which unifies these two phases of learning via the $\textit{skill discriminator}$, a pretraining model component often discarded during finetuning. Conventional approaches finetune pretrained agents directly at the policy level, often relying on expensive environment rollouts to empirically determine the optimal skill. However, often the most concise yet complete description of a task is the reward function itself, and skill learning methods learn an $\textit{intrinsic}$ reward function via the discriminator that corresponds to the skill policy. We propose to leverage the skill discriminator to $\textit{match}$ the intrinsic and downstream task rewards and determine the optimal skill for an unseen task without environment samples, consequently finetuning with greater sample-efficiency. Furthermore, we generalize IRM to sequence skills and solve more complex, long-horizon tasks. We demonstrate that IRM is competitive with previous skill selection methods on the Unsupervised Reinforcement Learning Benchmark and enables us to utilize pretrained skills far more effectively on challenging tabletop manipulation tasks.

****

TuneUp: A Training Strategy for Improving Generalization of Graph Neural Networks

Weihua Hu,Kaidi Cao,Kexin Huang,Edward W Huang,Karthik Subbian,Jure Leskovec

Despite many advances in Graph Neural Networks (GNNs), their training strategies simply focus on minimizing a loss over nodes in a graph. However, such simplistic training strategies may be sub-optimal as they neglect that certain nodes are much harder to make accurate predictions on than others. Here we present TuneUp, a curriculum learning strategy for better training GNNs. Crucially, TuneUp trains a GNN in two stages. The first stage aims to produce a strong base GNN. Such base GNNs tend to perform well on head nodes (nodes with large degrees) but less so on tail nodes (nodes with small degrees). So, the second stage of TuneUp specifically focuses on improving prediction on tail nodes. Concretely, TuneUp synthesizes many additional supervised tail node data by dropping edges from head nodes and reusing the supervision on the original head nodes. TuneUp then minimizes the loss over the synthetic tail nodes to finetune the base GNN. TuneUp is a general training strategy that can be used with any GNN architecture and any loss, making TuneUp applicable to a wide range of prediction tasks. Extensive evaluation of TuneUp on two GNN architectures, three types of prediction tasks, and both inductive and transductive settings shows that TuneUp significantly improves the performance of the base GNN on tail nodes, while often even improving the performance on head nodes, which together leads up to 58.5% relative improvement in GNN predictive performance. Moreover, TuneUp significantly outperforms its variants without the two-stage curriculum learning, existing graph data augmentation techniques, as well as other specialized methods for tail nodes.

****

A Scalable and Exact Gaussian Process Sampler via Kernel Packets

Haoyuan Chen,Rui Tuo

In view of the widespread use of Gaussian processes (GPs) in machine learning models, generating random sample paths of GPs is crucial for many machine learning applications. Sampling from a GP essentially requires generating high-dimensional Gaussian random vectors, which is computationally challenging if a direct method, such as the one based on Cholesky decomposition, is implemented. We develop a scalable algorithm to sample random realizations of the prior and the posterior of GP models with Matérn correlation functions. Unlike existing scalable sampling algorithms, the proposed approach draws samples from the theoretical distributions exactly. The algorithm exploits a novel structure called the kernel packets (KP), which gives an exact sparse representation of the dense covariance matrices. The proposed method is applicable for one-dimensional GPs, and multi-dimensional GPs under some conditions such as separable kernels with full grid designs. Via a series of experiments and comparisons with other recent works, we demonstrate the efficiency and accuracy of the proposed method.

****

Model ChangeLists: Characterizing Changes in ML Prediction APIs

Sabri Eyuboglu,Karan Goel,Arjun D Desai,Lingjiao Chen,Mathew Monfort,Matei Zaharia,Christopher Re,James Zou

Updates to Machine Learning as a Service (MLaaS) APIs may affect downstream systems that depend on their predictions. However, performance changes introduced by these updates are poorly documented by providers and seldom studied in the literature. As a result, users are left wondering: do model updates introduce subtle performance changes that could adversely affect my system? Ideally, users would have access to a detailed ChangeList specifying the slices of data where model performance has improved and degraded since the update. But, producing a ChangeList is challenging because it requires (1) discovering slices in the absence of detailed annotations or metadata, (2) accurately attributing coherent concepts to the discovered slices, and (3) communicating them to the user in a digestable manner. We introduce Mocha, an interactive framework for building, verifying and releasing ChangeLists that addresses these challenges. Using it, we perform a large-scale analysis of three real-world MLaaS API updates. We produce a ChangeList for each, identifying over 100 coherent data slices on which the model’s performance changed significantly. Notably, we find 63 instances where an update improves performance globally, but hurts performance on a coherent slice – a phenomenon not previously documented at scale in the literature. These findings underscore the importance of producing a detailed ChangeList when the model behind an API is updated.

****

Provably Auditing Ordinary Least Squares in Low Dimensions

Ankur Moitra,Dhruv Rohatgi

Auditing the stability of a machine learning model to small changes in the training procedure is critical for engendering trust in practical applications. For example, a model should not be overly sensitive to removing a small fraction of its training data. However, algorithmically validating this property seems computationally challenging, even for the simplest of models: Ordinary Least Squares (OLS) linear regression. Concretely, recent work defines the stability of a regression as the minimum number of samples that need to be removed so that rerunning the analysis overturns the conclusion (Broderick et al., 2020), specifically meaning that the sign of a particular coefficient of the OLS regressor changes. But the only known approach for estimating this metric, besides the obvious exponential-time algorithm, is a greedy heuristic that may produce severe overestimates and therefore cannot certify stability. We show that stability can be efficiently certified in the low-dimensional regime: when the number of covariates is a constant but the number of samples is large, there are polynomial-time algorithms for estimating (a fractional version of) stability, with provable approximation guarantees. Applying our algorithms to the Boston Housing dataset, we exhibit regression analyses where our estimator outperforms the greedy heuristic, and can successfully certify stability even in the regime where a constant fraction of the samples are dropped.

****

Exploring semantic information in disease: Simple Data Augmentation Techniques for Chinese Disease Normalization

Wenqian Cui,Xiangling Fu,Shaohui Liu,Xien Liu,Ji Wu

Disease is a core concept in the medical field, and the task of normalizing disease names is the basis of all disease-related tasks. However, due to the multi-axis and multi-grain nature of disease names, incorrect information is often injected and harms the performance when using general text data augmentation techniques. To address the above problem, we propose a set of data augmentation techniques that work together as an augmented training task for disease normalization, which is called Disease Data Augmentation (DDA). Our data augmentation methods are based on both the clinical disease corpus and standard disease corpus derived from ICD-10 coding. Extensive experiments are conducted to show the effectiveness of our proposed methods. The results demonstrate that our method can have up to 3\% performance gain compared to non-augmented counterparts, and they can work even better on smaller datasets.

****

Planning Goals for Exploration

Edward S. Hu,Richard Chang,Oleh Rybkin,Dinesh Jayaraman

Dropped into an unknown environment, what should an agent do to quickly learn about the environment and how to accomplish diverse tasks within it? We address this question within the goal-conditioned reinforcement learning paradigm, by identifying how the agent should set its goals at training time to maximize exploration. We propose "Planning Exploratory Goals" (PEG), a method that sets goals for each training episode to directly optimize an intrinsic exploration reward. PEG first chooses goal commands such that the agent's goal-conditioned policy, at its current level of training, will end up in states with high exploration potential. It then launches an exploration policy starting at those promising states. To enable this direct optimization, PEG learns world models and adapts sampling-based planning algorithms to "plan goal commands". In challenging simulated robotics environments including a multi-legged ant robot in a maze, and a robot arm on a cluttered tabletop, PEG exploration enables more efficient and effective training of goal-conditioned policies relative to baselines and ablations. Our ant successfully navigates a long maze, and the robot arm successfully builds a stack of three blocks upon command. Website: https://sites.google.com/view/exploratory-goals

****

Learning Sparse Group Models Through Boolean Relaxation

Yijie Wang,Yuan Zhou,Xiaoqing Huang,Kun Huang,Jie Zhang,Jianzhu Ma

We introduce an efficient algorithmic framework for learning sparse group models formulated as the natural convex relaxation of a cardinality-constrained program with Boolean variables. We provide theoretical techniques to characterize the equivalent condition when the relaxation achieves the exact integral optimal solution, as well as a rounding algorithm to produce a feasible integral solution once the optimal relaxation solution is fractional. We demonstrate the power of our equivalent condition by applying it to two ensembles of random problem instances that are challenging and popularly used in literature and prove that our method achieves exactness with overwhelming probability and nearly optimal sample complexity. Empirically, we use synthetic datasets to demonstrate that our proposed method significantly outperforms the state-of-the-art group sparse learning models in terms of individual and group support recovery when the number of samples is small. Furthermore, we show the out-performance of our method in cancer drug response prediction.

****

LVQ-VAE:End-to-end Hyperprior-based Variational Image Compression with Lattice Vector Quantization

Shinobu Kudo,Yukihiro Bandoh,Seishi Takamura,Masaki Kitahara

Image compression technology has become more important research topic. In recent years, learning-based methods have been extensively studied and variational autoencoder (VAE)-based methods using hyperprior-based context-adaptive entropy model have been reported to be comparable to the latest video coding standard H.266/VVC in terms of RD performance.

We think there is room for improvement in quantization process of latent features by adopting vector quantization (VQ). Many VAE-based methods use scalar quantization for latent features and do not exploit correlation between the features. Although there are methods that incorporate VQ into learning-based methods, to the best our knowledge, there are no studies that utilizes the hyperprior-based VAE with VQ because incorporating VQ into a hyperprior-based VAE makes it difficult to estimate the likelihood.

In this paper, we propose a new VAE-based image compression method using VQ based latent representation for hyperprior-based context-adaptive entropy model to improve the coding efficiency. The proposed method resolves problem faced by conventional VQ-based methods due to codebook size bloat by adopting Lattice VQ as the basis quantization method and achieves end-to-end optimization with hyperprior-based context-adaptive entropy model by approximating the likelihood calculation of latent feature vectors with high accuracy using Monte Carlo integration. Furthermore, in likelihood estimation, we model each latent feature vector with multivariate normal distribution including covariance matrix parameters, which improves the likelihood estimation accuracy and RD performance.

Experimental results show that the proposed method achieves a state-of-the-art RD performance exceeding existing learning-based methods and the latest video coding standard H.266/VVC by 18.0%.

****

Direct Embedding of Temporal Network Edges via Time-Decayed Line Graphs

Sudhanshu Chanpuriya,Ryan A. Rossi,Sungchul Kim,Tong Yu,Jane Hoffswell,Nedim Lipka,Shunan Guo,Cameron N Musco
Temporal networks model a variety of important phenomena involving timed interactions between entities. Existing methods for machine learning on temporal networks generally exhibit at least one of two limitations. First, many methods assume time to be discretized, so if the time data is continuous, the user must determine the discretization and discard precise time information. Second, edge representations can only be calculated indirectly from the nodes, which may be suboptimal for tasks like edge classification. We present a simple method that avoids both shortcomings: construct the line graph of the network, which includes a node for each interaction, and weigh the edges of this graph based on the difference in time between interactions. From this derived graph, edge representations for the original network can be computed with efficient classical methods. The simplicity of this approach facilitates explicit theoretical analysis: we can constructively show the effectiveness of our method's representations for a natural synthetic model of temporal networks. Empirical results on real-world networks demonstrate our method's efficacy and efficiency on both link classification and prediction.
****
Neural DAG Scheduling via One-Shot Priority Sampling
Wonseok Jeon,Mukul Gagrani,Burak Bartan,Weiliang Will Zeng,Harris Teague,Piero Zappi,Christopher Lott
We consider the problem of scheduling operations/nodes, the dependency among which is characterized by a Directed Acyclic Graph (DAG). Due to its NP-hard nature, heuristic algorithms were traditionally used to acquire reasonably good solutions, and more recent works have proposed Machine Learning (ML) heuristics that can generalize to unseen graphs and outperform the non-ML heuristics. However, it is computationally costly to generate solutions using existing ML schedulers since they adopt the episodic reinforcement learning framework that necessitates multi-round neural network processing. We propose a novel ML scheduler that uses a one-shot neural network encoder to sample node priorities which are converted by list scheduling to the final schedules. Since the one-shot encoder can efficiently sample the priorities in parallel, our algorithm runs significantly faster than existing ML baselines and has comparable run time with the fast traditional heuristics. We empirically show that our algorithm generates better schedules than both non-neural and neural baselines across various real-world and synthetic scheduling tasks.
****
Efficiently Computing Nash Equilibria in Adversarial Team Markov Games
Fivos Kalogiannis,Ioannis Anagnostides,Ioannis Panageas,Emmanouil-Vasileios Vlatakis-Gkaragkounis,Vaggos Chatziafratis,Stelios Andrew Stavroulakis

Computing Nash equilibrium policies is a central problem in multi-agent reinforcement learning that has received extensive attention both in theory and in practice. However, in light of computational intractability barriers in general-sum games, provable guarantees have been thus far either limited to fully competitive or cooperative scenarios or impose strong assumptions that are difficult to meet in most practical applications.

In this work, we depart from those prior results by investigating infinite-horizon \emph{adversarial team Markov games}, a natural and well-motivated class of games in which a team of identically-interested players---in the absence of any explicit coordination or communication---is competing against an adversarial player. This setting allows for a unifying treatment of zero-sum Markov games and Markov potential games, and serves as a step to model more realistic strategic interactions that feature both competing and cooperative interests. Our main contribution is the first algorithm for computing stationary ϵ-approximate Nash equilibria in adversarial team Markov games with computational complexity that is polynomial in all the natural parameters of the game, as well as $1/\epsilon$.

The proposed algorithm is based on performing independent policy gradient steps for each player in the team, in tandem with best responses from the side of the adversary; in turn, the policy for the adversary is then obtained by solving a carefully constructed linear program. Our analysis leverages non-standard techniques to establish the KKT optimality conditions for a nonlinear program with nonconvex constraints, thereby leading to a natural interpretation of the induced Lagrange multipliers.

****

Meta Temporal Point Processes

Wonho Bae,Mohamed Osama Ahmed,Frederick Tung,Gabriel L. Oliveira

A temporal point process (TPP) is a stochastic process where its realization is a sequence of discrete events in time. Recent work in TPPs model the process using a neural network in a supervised learning framework, where a training set is a collection of all the sequences. In this work, we propose to train TPPs in a meta learning framework, where each sequence is treated as a different task, via a novel framing of TPPs as neural processes (NPs). We introduce context sets to model TPPs as an instantiation of NPs. Motivated by attentive NP, we also introduce local history matching to help learn more informative features. We demonstrate the potential of the proposed method on popular public benchmark datasets and tasks, and compare with state-of-the-art TPP methods.

****

EmbedDistill: A geometric knowledge distillation for information retrieval

Seungyeon Kim,Ankit Singh Rawat,Manzil Zaheer,Sadeep Jayasumana,Veeranjaneyulu Sadhanala,Wittawat Jitkrittum,Aditya Krishna Menon,Rob Fergus,Sanjiv Kumar

Large neural models (such as Transformers) achieve state-of-the-art performance for information retrieval. In this paper, we aim to improve distillation methods that pave the way for the deployment of such models in practice. The proposed distillation approach supports both retrieval and re-ranking stages and crucially leverages the relative geometry among queries and documents learned by the large teacher model. It goes beyond existing distillation methods in the information retrieval literature, which simply rely on the teacher's scalar scores over the training data, on two fronts: providing stronger signals about local geometry via embedding matching and attaining better coverage of data manifold globally via query generation. Embedding matching provides a stronger signal to align the representations of the teacher and student models. At the same time, query generation explores the data manifold to reduce the discrepancies between the student and teacher where the training data is sparse. Our distillation approach is theoretically justified and applies to both dual encoder (DE) and cross-encoder (CE) models. Furthermore, for distilling a CE model to a DE model via embedding matching, we propose a novel dual pooling-based scorer for the CE model that facilitates a more distillation-friendly embedding geometry, especially for DE student models.

****

Graph Neural Network-Inspired Kernels for Gaussian Processes in Semi-Supervised Learning

Zehao Niu,Mihai Anitescu,Jie Chen

Gaussian processes (GPs) are an attractive class of machine learning models because of their simplicity and flexibility as building blocks of more complex Bayesian models. Meanwhile, graph neural networks (GNNs) emerged recently as a promising class of models for graph-structured data in semi-supervised learning and beyond. Their competitive performance is often attributed to a proper capturing of the graph inductive bias. In this work, we introduce this inductive bias into GPs to improve their predictive performance for graph-structured data. We show that a prominent example of GNNs, the graph convolutional network, is equivalent to some GP when its layers are infinitely wide; and we analyze the kernel universality and the limiting behavior in depth. We further present a programmable procedure to compose covariance kernels inspired by this equivalence and derive example kernels corresponding to several interesting members of the GNN family. We also propose a computationally efficient approximation of the covariance matrix for scalable posterior inference with large-scale data. We demonstrate that these

graph-based kernels lead to competitive classification and regression performance, as well as advantages in computation time, compared with the respective GNNs.
****
Deconstructing Distributions: A Pointwise Framework of Learning
Gal Kaplun,Nikhil Ghosh,Saurabh Garg,Boaz Barak,Preetum Nakkiran
In machine learning, we traditionally evaluate the performance of a single model, averaged over a collection of test inputs. In this work, we propose a new approach: we measure the performance of a collection of models when evaluated at *single input point*. Specifically, we study a point's *profile*: the relationship between models' average performance on the test distribution and their pointwise performance on this individual point. We find that profiles can yield new insights into the structure of both models and data---in and out-of-distribution. For example, we empirically show that real data distributions consist of points with qualitatively different profiles. On one hand, there are ``compatible'' points with strong correlation between the pointwise and average performance. On the other hand, there are points with weak and even *negative* correlation: cases where improving overall model accuracy actually *hurts* performance on these inputs. As an application, we use profiles to construct a dataset we call CIFAR-10-NEG: a subset of CINIC-10 such that for standard models, accuracy on CIFAR-10-NEG is *negatively correlated* with CIFAR-10 accuracy. Illustrating for the first time an OOD dataset that completely inverts ``accuracy-on-the-line'' (Miller et al., 2021).
****
Logical view on fairness of a binary classification task
Serge Berger
Ethical, Interpretable/Explainable, and Responsible AI are an active area of research and important social initiative.

We prove that, with no regards to data, fairness and trustworthiness are algorithmically undecidable for a basic machine learning task, the binary classification. Therefore, even the approach based on not only improving but fully solving the three usually assumed issues -- the insufficient quality of measurements, the complex consequences of (mis)measurements, and the limits of existing social theories -- is only heuristics. We show that, effectively, the fairness of a classifier is not even a (version of bias-variance) trade-off since it is a logical phenomenon.
Namely, we reveal a language L and an $L-$theory T for binary classification task such that the very notion of loss is not expressible in the first-order logic formula in L.
****
Revisiting Instance-Reweighted Adversarial Training
Hiroki Adachi,Tsubasa Hirakawa,Takayoshi Yamashita,Hironobu Fujiyoshi
Instance-reweighted adversarial training (IRAT) is a type of adversarial training that assigns large weights to high-importance examples and then minimizes the weighted loss. The importance often uses the margins between decision boundaries and each example. In particular, IRAT can alleviate robust overfitting and obtain excellent robustness by computing margins with an estimated probability. However, previous works implicitly dealt with binary classification even in the multi-class cases, because they computed margins with only the true class and the most confusing class. The computed margins can become equal even with different true probability examples, because of the complex decision boundaries in multi-class classification. In this paper, first, we clarify the above problem with a specific example. Then, we propose \textit{margin reweighting}, which can transform the previous margins into appropriate representations for multi-class classification by leveraging the relations between the most confusing class and other classes. Experimental results on the CIFAR-10/100 datasets demonstrate that the proposed method is effective in boosting the robustness against several attacks as compared to the previous methods.
****

Diffusion Models for Causal Discovery via Topological Ordering

Pedro Sanchez,Xiao Liu,Alison Q O'Neil,Sotirios A. Tsaftaris

Discovering causal relations from observational data becomes possible with additional assumptions such as considering the functional relations to be constrained as nonlinear with additive noise (ANM). Even with strong assumptions, causal discovery involves an expensive search problem over the space of directed acyclic graphs (DAGs). \emph{Topological ordering} approaches reduce the optimisation space of causal discovery by searching over a permutation rather than graph space. For ANMs, the \emph{Hessian} of the data log-likelihood can be used for finding leaf nodes in a causal graph, allowing its topological ordering. However, existing computational methods for obtaining the Hessian still do not scale as the number of variables and the number of samples are increased. Therefore, inspired by recent innovations in diffusion probabilistic models (DPMs), we propose \emph{DiffAN}, a topological ordering algorithm that leverages DPMs for learning a Hessian function. We introduce theory for updating the learned Hessian without re-training the neural network, and we show that computing with a subset of samples gives an accurate approximation of the ordering, which allows scaling to datasets with more samples and variables. We show empirically that our method scales exceptionally well to datasets with up to 500 nodes and up to 10^5 samples while still performing on par over small datasets with state-of-the-art causal discovery methods.

Implementation is available at \url{https://github.com/vios-s/DiffAN} .

****

Scalable and Equivariant Spherical CNNs by Discrete-Continuous (DISCO) Convolutions

Jeremy Ocampo,Matthew Alexander Price,Jason McEwen

No existing spherical convolutional neural network (CNN) framework is both computationally scalable and rotationally equivariant. Continuous approaches capture rotational equivariance but are often prohibitively computationally demanding. Discrete approaches offer more favorable computational performance but at the cost of equivariance. We develop a hybrid discrete-continuous (DISCO) group convolution that is simultaneously equivariant and computationally scalable to high-resolution. While our framework can be applied to any compact group, we specialize to the sphere. Our DISCO spherical convolutions exhibit $\text{SO}(3)$ rotational equivariance, where $\text{SO}(n)$ is the special orthogonal group representing rotations in n-dimensions. When restricting rotations of the convolution to the quotient space $\text{SO}(3)/\text{SO}(2)$ for further computational enhancements, we recover a form of asymptotic $\text{SO}(3)$ rotational equivariance. Through a sparse tensor implementation we achieve linear scaling in number of pixels on the sphere for both computational cost and memory usage. For 4k spherical images we realize a saving of 10^9 in computational cost and 10^4 in memory usage when compared to the most efficient alternative equivariant spherical convolution. We apply the DISCO spherical CNN framework to a number of benchmark dense-prediction problems on the sphere, such as semantic segmentation and depth estimation, on all of which we achieve the state-of-the-art performance.

****

Towards Solving Industrial Sequential Decision-making Tasks under Near-predictable Dynamics via Reinforcement Learning: an Implicit Corrective Value Estimation Approach

Jianyong Yuan,Jiayi Zhang,Junchi Yan

Learning to plan and schedule is receiving increasing attention for industrial decision-making tasks for its potential for outperforming heuristics, especially under dynamic uncertainty, as well as its efficiency in problem-solving, especially with the adoption of neural networks and the behind GPU computing. Naturally, reinforcement learning (RL) with the Markov decision process (MDP) becomes a popular paradigm. Rather than handling the near-stationary environments like Atari games or the opposite case for open world dynamics with high uncertainty. In this paper, we aim to devise a tailored RL-based approach for the setting in the between: the near-predictable dynamics which often hold in many industrial applications, e.g., elevator scheduling and bin packing, as empirical case studies te sted in this paper. We formulate a two-stage MDP by decoupling the data dynamics from the industrial environment. Specifically, we design a bi-critic framework for estimating the state value in stages according to the two-stage MDP.

****

Graph Convolutional Normalizing Flows for Semi-Supervised Classification and Clustering

Tianchun Wang,Farzaneh Mirzazadeh,Xiang Zhang,Jie Chen

Graph neural networks (GNNs) are \emph{discriminative models} that directly model the class posterior $p(y|\mathbf{x})$ for semi-supervised classification of graph data. While being effective for prediction, as a representation learning approach, the node representations extracted from a GNN often miss useful information for effective clustering, because that is not necessary for a good classification. In this work, we replace a GNN layer by a combination of graph convolutions and normalizing flows under a Gaussian mixture representation space, which allows us to build a \emph{generative model} that models both the class conditional likelihood $p(\mathbf{x}|y)$ and the class prior $p(y)$. The resulting neural network, GC-Flow, enjoys two benefits: it not only maintains the predictive power because of the retention of graph convolutions, but also produces high-quality clusters in the representation space, due to the structuring of the representation as a mixture of Gaussians. We demonstrate these benefits on a variety of benchmark data sets. Moreover, we show that additional parameterization, such as that on the adjacency matrix used for graph convolutions, yields additional improvement in classification and clustering.

****

Weakly Supervised Explainable Phrasal Reasoning with Neural Fuzzy Logic

Zijun Wu,Zi Xuan Zhang,Atharva Naik,Zhijian Mei,Mauajama Firdaus,Lili Mou

Natural language inference (NLI) aims to determine the logical relationship between two sentences, such as Entailment, Contradiction, and Neutral. In recent years, deep learning models have become a prevailing approach to NLI, but they lack interpretability and explainability. In this work, we address the explainability of NLI by weakly supervised logical reasoning, and propose an Explainable Phrasal Reasoning (EPR) approach. Our model first detects phrases as the semantic unit and aligns corresponding phrases in the two sentences. Then, the model predicts the NLI label for the aligned phrases, and induces the sentence label by fuzzy logic formulas. Our EPR is almost everywhere differentiable and thus the system can be trained end to end. In this way, we are able to provide explicit explanations of phrasal logical relationships in a weakly supervised manner. We further show that such reasoning results help textual explanation generation.

****

Simplified State Space Layers for Sequence Modeling

Jimmy T.H. Smith,Andrew Warrington,Scott Linderman

Models using structured state space sequence (S4) layers have achieved state-of-the-art performance on long-range sequence modeling tasks. An S4 layer combines linear state space models (SSMs), the HiPPO framework, and deep learning to achieve high performance. We build on the design of the S4 layer and introduce a new state space layer, the S5 layer. Whereas an S4 layer uses many independent single-input, single-output SSMs, the S5 layer uses one multi-input, multi-output SSM. We establish a connection between S5 and S4, and use this to develop the initialization and parameterization used by the S5 model. The result is a state space layer that can leverage efficient and widely implemented parallel scans, allowing S5 to match the computational efficiency of S4, while also achieving state-of-the-art performance on several long-range sequence modeling tasks. S5 averages 87.4% on the long range arena benchmark, and 98.5% on the most difficult Path-X task.

****

Learning Listwise Domain-Invariant Representations for Ranking

Ruicheng Xian,Honglei Zhuang,Zhen Qin,Hamed Zamani,Jing Lu,Ji Ma,Kai Hui,Han Zhao,Xuanhui Wang,Michael Bendersky

Domain adaptation aims to transfer models trained on data-rich domains to low-re

source ones, for which a popular method is invariant representation learning. While they have been studied extensively for classification and regression problems, how they would apply to ranking problems, where the metrics and data follow a list structure, is not well understood. Theoretically, we establish a generalization bound for ranking problems under metrics including MRR and NDCG, leading to a method based on learning listwise invariant feature representations. The main novelty of our results is that they are tailored to the listwise approach of learning to rank: the invariant representations our method learns are for each list of items as a whole, instead of the individual items they contain. Our method is evaluated on the passage reranking task, where we adapt neural text rankers trained on a general domain to various specialized domains.
****
DCI-ES: An Extended Disentanglement Framework with Connections to Identifiability
Cian Eastwood,Andrei Liviu Nicolicioiu,Julius Von Kügelgen,Armin Keki■,Frederik Träuble,Andrea Dittadi,Bernhard Schölkopf
In representation learning, a common approach is to seek representations which disentangle the underlying factors of variation. Eastwood & Williams (2018) proposed three metrics for quantifying the quality of such disentangled representations: disentanglement (D), completeness (C) and informativeness (I). In this work, we first connect this DCI framework to two common notions of linear and nonlinear identifiability, thereby establishing a formal link between disentanglement and the closely-related field of independent component analysis. We then propose an extended DCI-ES framework with two new measures of representation quality—explicitness (E) and size (S)—and point out how D and C can be computed for black-box predictors. Our main idea is that the functional capacity required to use a representation is an important but thus-far neglected aspect of representation quality, which we quantify using explicitness or ease-of-use (E). We illustrate the relevance of our extensions on the MPI3D and Cars3D datasets.
****
Eigenvalue Initialisation and Regularisation for Koopman Autoencoders
Jack William Miller,Charles O'Neill,Navid C Constantinou,Omri Azencot
Regularising the parameter matrices of neural networks is ubiquitous in training deep models. Typical regularisation approaches suggest initialising weights using small random values, and to penalise weights to promote sparsity. However, these widely used techniques may be less effective in certain scenarios. Here, we study the Koopman autoencoder model which includes an encoder, a Koopman operator layer, and a decoder. These models have been designed and dedicated to tackle physics-related problems with interpretable dynamics and an ability to incorporate physics-related constraints. However, the majority of existing work employs standard regularisation practices. In our work, we take a step toward augmenting Koopman autoencoders with initialisation and penalty schemes tailored for physics-related settings. Specifically, we propose the "eigeninit" initialisation scheme that samples initial Koopman operators from specific eigenvalue distributions. In addition, we suggest the "eigenloss" penalty scheme that penalises the eigenvalues of the Koopman operator during training. We demonstrate the utility of these schemes on two synthetic data sets: a driven pendulum and flow past a cylinder; and two real-world problems: ocean surface temperatures and cyclone wind fields. We find on these datasets that eigenloss and eigeninit improves the convergence rate by a factor of 2 to 5, and that they reduce the cumulative long-term prediction error by up to a factor of 2.5. Such a finding points to the utility of incorporating similar schemes as an inductive bias in other physics-related deep learning approaches.
****
Learning from Labeled Images and Unlabeled Videos for Video Segmentation
Cristina Mata,Michael S Ryoo
Performance on video object segmentation still lags behind that of image segmentation due to a paucity of labeled videos. Annotations are time-consuming and laborious to collect, and may not be feasibly obtained in certain situations. However there is a growing amount of freely available unlabeled video data which has

spurred interest in unsupervised video representation learning. In this work we focus on the setting in which there is no/little access to labeled videos for video object segmentation. To this end we leverage large-scale image segmentation datasets and adversarial learning to train 2D/3D networks for video object segmentation. We first motivate the treatment of images and videos as two separate domains by analyzing the performance gap of an image segmentation network trained on images and applied to videos. Through studies using several image and video segmentation datasets, we show how an adversarial loss placed at various locations within the network can make feature representations invariant to these domains and improve the performance when the network has access to only labeled images and unlabeled videos. To prevent the loss of discriminative semantic class information we apply our adversarial loss within clusters of features and show this boosts our method's performance within Transformer-based models.

****

MeshDiffusion: Score-based Generative 3D Mesh Modeling
Zhen Liu,Yao Feng,Michael J. Black,Derek Nowrouzezahrai,Liam Paull,Weiyang Liu
We consider the task of generating realistic 3D shapes, which is useful for a variety of applications such as automatic scene generation and physical simulation. Compared to other 3D representations like voxels and point clouds, meshes are more desirable in practice, because (1) they enable easy and arbitrary manipulation of shapes for relighting and simulation, and (2) they can fully leverage the power of modern graphics pipelines which are mostly optimized for meshes. Previous scalable methods for generating meshes typically rely on sub-optimal post-processing, and they tend to produce overly-smooth or noisy surfaces without fine-grained geometric details. To overcome these shortcomings, we take advantage of the graph structure of meshes and use a simple yet very effective generative modeling method to generate 3D meshes. Specifically, we represent meshes with deformable tetrahedral grids, and then train a diffusion model on this direct parameterization. We demonstrate the effectiveness of our model on multiple generative tasks.

****

Faster federated optimization under second-order similarity
Ahmed Khaled,Chi Jin
Federated learning (FL) is a subfield of machine learning where multiple clients try to collaboratively learn a model over a network under communication constraints. We consider finite-sum federated optimization under a second-order function similarity condition and strong convexity, and propose two new algorithms: SVRP and Catalyzed SVRP. This second-order similarity condition has grown popular recently, and is satisfied in many applications including distributed statistical learning and differentially private empirical risk minimization. The first algorithm, SVRP, combines approximate stochastic proximal point evaluations, client sampling, and variance reduction. We show that SVRP is communication efficient and achieves superior performance to many existing algorithms when function similarity is high enough. Our second algorithm, Catalyzed SVRP, is a Catalyst-accelerated variant of SVRP that achieves even better performance and uniformly improves upon existing algorithms for federated optimization under second-order similarity and strong convexity. In the course of analyzing these algorithms, we provide a new analysis of the Stochastic Proximal Point Method (SPPM) that might be of independent interest. Our analysis of SPPM is simple, allows for approximate proximal point evaluations, does not require any smoothness assumptions, and shows a clear benefit in communication complexity over ordinary distributed stochastic gradient descent.

****

A Quasistatic Derivation of Optimization Algorithms' Exploration on Minima Manifolds
Chao Ma,Daniel Kunin,Lexing Ying
A quasistatic approach is proposed to derive the optimization algorithms' effective dynamics on the manifold of minima when the iterator oscillates around the manifold. Compared with existing strict analysis, our derivation method is simple and intuitive, has wide applicability, and produces easy-to-interpret results.

As examples, we derive the manifold dynamics for SGD, SGD with momentum (SGDm) and Adam with different noise covariances, and justify the closeness of the derived manifold dynamics with the true dynamics through numerical experiments. We then use minima manifold dynamics to study and compare the properties of optimization algorithms. For SGDm, we show that scaling up learning rate and batch size simultaneously accelerates exploration without affecting generalization, which confirms a benefit of large batch training. For Adam, we show that the speed of its manifold dynamics changes with the direction of the manifold, because Adam is not rotationally invariant. This may cause slow exploration in high dimensional parameter spaces.

****

Mutual Partial Label Learning with Competitive Label Noise

Yan Yan,Yuhong Guo

Partial label learning (PLL) is an important weakly supervised learning problem, where each training instance is associated with a set of candidate labels that include both the true label and additional noisy labels. Most existing PLL methods assume the candidate noisy labels are randomly chosen, which hardly holds in real-world learning scenarios. In this paper, we consider a more realistic PLL scenario with competitive label noise that is more difficult to distinguish from the true label than the random label noise. We propose a novel Mutual Learning based PLL approach named ML-PLL to address this challenging problem. ML-PLL learns a prediction network based classifier and a class-prototype based classifier cooperatively through interactive mutual learning and label correction. Moreover, we use a transformation network to model the association relationships between the true label and candidate labels, and learn it together with the prediction network to match the observed candidate labels in the training data and enhance label correction. Extensive experiments are conducted on several benchmark PLL datasets, and the proposed ML-PLL approach demonstrates state-of-the-art performance for partial label learning.

****

The Graph Learning Attention Mechanism: Learnable Sparsification Without Heuristics

Mattson Thieme,Yada Zhu,Han Liu

Graph Neural Networks (GNNs) are local aggregators that derive their expressive power from their sensitivity to network structure. However, this sensitivity comes at a cost: noisy edges degrade performance. In response, many GNNs include edge-weighting mechanisms that scale the contribution of each edge in the aggregation step. However, to account for neighborhoods of varying size, node-embedding mechanisms must normalize these edge-weights across each neighborhood. As such, the impact of noisy edges cannot be eliminated without removing those edges altogether. Motivated by this issue, we introduce the Graph Learning Attention Mechanism (GLAM): a drop-in, differentiable structure learning layer for GNNs that separates the distinct tasks of structure learning and node embedding. In contrast to existing graph learning approaches, GLAM does not require the addition of exogenous structural regularizers or edge-selection heuristics to learn optimal graph structures. In experiments on citation and co-purchase datasets, we demonstrate that our approach can match state of the art semi-supervised node classification accuracies while inducing an order of magnitude greater sparsity than existing graph learning methods.

****

Partial Label Unsupervised Domain Adaptation with Class-Prototype Alignment

Yan Yan,Yuhong Guo

Partial label learning (PLL) tackles the problem where each instance is associated with a set of candidate labels, only one of which is the ground-truth label. Most existing PLL approaches assume that both the training and test sets share an identical data distribution. However, this assumption does not hold in many real-world scenarios where the training and test data come from different distributions. In this paper, we formalize this learning scenario as a new problem called partial label unsupervised domain adaptation (PLUDA). To address this challenging PLUDA problem, we propose a novel Prototype Alignment based PLUDA method nam ed PAPLUDA, which dynamically refines the pseudo-labels of instances from both the source and target domains by consulting the outputs of a teacher-student model in a moving-average manner, and bridges the cross-domain discrepancy through inter-domain class-prototype alignment. In addition, a teacher-student model based contrastive regularization is deployed to enhance prediction stability and hence improve the class-prototypes in both domains for PLUDA. Comprehensive experimental results demonstrate that PAPLUDA achieves state-of-the-art performance on the widely used benchmark datasets.

****

Why Self Attention is Natural for Sequence-to-Sequence Problems? A Perspective from Symmetries

Chao Ma,Lexing Ying

In this paper, we show that structures similar to self-attention are natural to learn many sequence-to-sequence problems from the perspective of symmetry. Inspired by language processing applications, we study the orthogonal equivariance of {\it seq2seq functions with knowledge}, which are functions taking two inputs---an input sequence and a ``knowledge''---and outputting another sequence.

The knowledge consists of a set of vectors in the same embedding space as the input sequence, containing the information of the language used to process the input sequence. We show that orthogonal equivariance in the embedding space is natural for seq2seq functions with knowledge, and under such equivariance the function must take the form close to the self-attention. This shows that network structures similar to self-attention are the right structures to represent the target function of many seq2seq problems. The representation can be further refined if a ``finite information principle'' is considered, or a permutation equivariance holds for the elements of the input sequence.

****

simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing

Zitao Liu,Qiongqiong Liu,Jiahao Chen,Shuyan Huang,Weiqi Luo

Knowledge tracing (KT) is the problem of predicting students' future performance based on their historical interactions with intelligent tutoring systems. Recently, many works present lots of special methods for applying deep neural networks to KT from different perspectives like model architecture, adversarial augmentation and etc., which make the overall algorithm and system become more and more complex. Furthermore, due to the lack of standardized evaluation protocol \citep{liu2022pykt}, there is no widely agreed KT baselines and published experimental comparisons become inconsistent and self-contradictory, i.e., the reported AUC scores of DKT on ASSISTments2009 range from 0.721 to 0.821 \citep{minn2018deep,yeung2018addressing}. Therefore, in this paper, we provide a strong but simple baseline method to deal with the KT task named \textsc{simpleKT}. Inspired by the Rasch model in psychometrics, we explicitly model question-specific variations to capture the individual differences among questions covering the same set of knowledge components that are a generalization of terms of concepts or skills needed for learners to accomplish steps in a task or a problem. Furthermore, instead of using sophisticated representations to capture student forgetting behaviors, we use the ordinary dot-product attention function to extract the time-aware information embedded in the student learning interactions. Extensive experiments show that such a simple baseline is able to always rank top 3 in terms of AUC scores and achieve 57 wins, 3 ties and 16 loss against 12 DLKT baseline methods on 7 public datasets of different domains. We believe this work serves as a strong baseline for future KT research. Code is available at \url{https://github.com/pykt-team/pykt-toolkit}\footnote{We merged our model to the \textsc{pyKT} benchmark at \url{https://pykt.org/}.}.

****

Exp-α: Beyond Proportional Aggregation in Federated Learning

Junjiao Tian,Xiaoliang Dai,Chih-Yao Ma,Zecheng He,Yen-Cheng Liu,Sayan Ghosh,Peter Vajda,Anqi Wu,Zsolt Kira

Federated Learning (FL) is a distributed learning paradigm, which computes gradients of a model locally on different clients and aggregates the updates to construct a new model collectively. Typically, the updates from local clients are agg regated with weights proportional to the size of clients' local datasets. In practice, clients have different local datasets suffering from data heterogeneity, such as imbalance. Although proportional aggregation still theoretically converges to the global optimum, it is provably slower when non-IID data is present (under convexity assumptions), the effect of which is exacerbated in practice. We posit that this analysis ignores convergence rate, which is especially important under such settings in the more realistic non-convex real world. To account for this, we analyze a generic and time-varying aggregation strategy to reveal a surprising trade-off between convergence rate and convergence error under convexity assumptions. Inspired by the theory, we propose a new aggregation strategy, Exp-α, which weights clients differently based on their severity of data heterogeneity. It achieves stronger convergence rates at the theoretical cost of a non-vanishing convergence error. Through a series of controlled experiments, we empirically demonstrate the superior convergence behavior (both in terms of rate and, in practice, even error) of the proposed aggregation on three types of data heterogeneity: imbalance, label-flipping, and domain shift when combined with existing FL algorithms. For example, on our imbalance benchmark, Exp-α, combined with FedAvg, achieves a relative 12% increase in convergence rate and a relative 3% reduction in error across four FL communication settings.

****

Learning Efficient Hybrid Particle-continuum Representations of Non-equilibrium N-body Systems

Tailin Wu,Michael Sun,H.G. Jason Chou,Pranay Samala,Sithipont Cholsaipant,Sophia Kivelson,Jacqueline Yau,Zhitao Ying,E. Paulo Alves,Jure Leskovec,Frederico Fiuza

An important class of multi-scale, non-equilibrium, N-body physical systems deals with an interplay between particle and continuum phenomena. These include hypersonic flow and plasma dynamics, materials science, and astrophysics. Hybrid solvers that combine particle and continuum representations could provide an efficient framework to model these systems. However, the coupling between these two representations has been a key challenge, which is often limited to inaccurate or incomplete prescriptions. In this work, we introduce a method for Learning Hybrid Particle-Continuum (LHPC) models from the data of first-principles particle simulations. LHPC analyzes the local velocity-space particle distribution function and separates it into near-equilibrium (thermal) and far-from-equilibrium (non-thermal) components. The most computationally-intensive particle solver is used to advance the non-thermal particles, whereas a neural network solver is used to efficiently advance the thermal component using a continuum representation. Most importantly, an additional neural network learns the particle-continuum coupling: the dynamical exchange of mass, momentum, and energy between the particle and continuum representations. Training of the different neural network components is done in an integrated manner to ensure global consistency and stability of the LHPC model. We demonstrate our method in an intense laser-plasma interaction problem involving highly nonlinear, far-from-equilibrium dynamics associated with the coupling between electromagnetic fields and multiple particle species. More efficient modeling of these interactions is critical for the design and optimization of compact accelerators for material science and medical applications. Our method achieves an important balance between accuracy and speed: LHPC is 8 times faster than a classical particle solver and achieves up to 6.8-fold reduction of long-term prediction error for key quantities of interest compared to deep-learning baselines using uniform representations.

****

Neural Network Approximation of Lipschitz Functions in High Dimensions with Applications to Inverse Problems

Santhosh Karnik,Rongrong Wang,Mark Iwen

The remarkable successes of neural networks in a huge variety of inverse problems have fueled their adoption in disciplines ranging from medical imaging to seismic analysis over the past decade. However, the high dimensionality of such inverse problems has simultaneously left current theory, which predicts that networks should scale exponentially in the dimension of the problem, unable to explain why the seemingly small networks used in these settings work as well as they do in practice. To reduce this gap between theory and practice, a general method for bounding the complexity required for a neural network to approximate a Lipschitz function on a high-dimensional set with a low-complexity structure is provided herein. The approach is based on the observation that the existence of a linear Johnson-Lindenstrauss embedding $\mathbf{A} \in \mathbb{R}^{d \times D}$ of a given high-dimensional set $\mathcal{S} \subset \mathbb{R}^D$ into a low dimensional cube $[-M,M]^d$ implies that for any Lipschitz function $f : \mathcal{S} \to \mathbb{R}^p$, there exists a Lipschitz function $g : [-M,M]^d \to \mathbb{R}^p$ such that $g(\mathbf{A}\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$. Hence, if one has a neural network which approximates $g : [-M,M]^d \to \mathbb{R}^p$, then a layer can be added which implements the JL embedding $\mathbf{A}$ to obtain a neural network which approximates $f : \mathcal{S} \to \mathbb{R}^p$. By pairing JL embedding results along with results on approximation of Lipschitz functions by neural networks, one then obtains results which bound the complexity required for a neural network to approximate Lipschitz functions on high dimensional sets. The end result is a general theoretical framework which can then be used to better explain the observed empirical successes of smaller networks in a wider variety of inverse problems than current theory allows.

****

Weighted Ensemble Self-Supervised Learning

Yangjun Ruan,Saurabh Singh,Warren Richard Morningstar,Alexander A Alemi,Sergey Ioffe,Ian Fischer,Joshua V. Dillon

Ensembling has proven to be a powerful technique for boosting model performance, uncertainty estimation, and robustness in supervised learning. Advances in self-supervised learning (SSL) enable leveraging large unlabeled corpora for state-of-the-art few-shot and supervised learning performance. In this paper, we explore how ensemble methods can improve recent SSL techniques by developing a framework that permits data-dependent weighted cross-entropy losses. We refrain from ensembling the representation backbone; this choice yields an efficient ensemble method that incurs a small training cost and requires no architectural changes or computational overhead to downstream evaluation. The effectiveness of our method is demonstrated with two state-of-the-art SSL methods, DINO (Caron et al., 2021) and MSN (Assran et al., 2022). Our method outperforms both in multiple evaluation metrics on ImageNet-1K, particularly in the few-shot setting. We explore several weighting schemes and find that those which increase the diversity of ensemble heads lead to better downstream evaluation results. Thorough experiments yield improved prior art baselines which our method still surpasses; e.g., our overall improvement with MSN ViT-B/16 is 3.9 p.p. for 1-shot learning.

****

Partially Observable RL with B-Stability: Unified Structural Condition and Sharp Sample-Efficient Algorithms

Fan Chen,Yu Bai,Song Mei

Partial Observability---where agents can only observe partial information about the true underlying state of the system---is ubiquitous in real-world applications of Reinforcement Learning (RL). Theoretically, learning a near-optimal policy under partial observability is known to be hard in the worst case due to an exponential sample complexity lower bound. Recent work has identified several tractable subclasses that are learnable with polynomial samples, such as Partially Observable Markov Decision Processes (POMDPs) with certain revealing or decodability conditions. However, this line of research is still in its infancy, where (1) unified structural conditions enabling sample-efficient learning are lacking; (2) existing sample complexities for known tractable subclasses are far from sharp; and (3) fewer sample-efficient algorithms are available than in fully observable RL.

This paper advances all three aspects above for Partially Observable RL in the general setting of Predictive State Representations (PSRs). First, we propose a natural and unified structural condition for PSRs called \emph{B-stability}. B-stable PSRs encompasses the vast majority of known tractable subclasses such as we akly revealing POMDPs, low-rank future-sufficient POMDPs, decodable POMDPs, and regular PSRs. Next, we show that any B-stable PSR can be learned with polynomial samples in relevant problem parameters. When instantiated in the aforementioned subclasses, our sample complexities improve substantially over the current best ones. Finally, our results are achieved by three algorithms simultaneously: Optimistic Maximum Likelihood Estimation, Estimation-to-Decisions, and Model-Based Optimistic Posterior Sampling. The latter two algorithms are new for sample-efficient learning of POMDPs/PSRs.

We additionally design a variant of the Estimation-to-Decisions algorithm to perform sample-efficient \emph{all-policy model estimation} for B-stable PSRs, which also yields guarantees for reward-free learning as an implication.

****

Bias Amplification Improves Worst-Group Accuracy without Group Information

Gaotang Li,Jiarui Liu,Wei Hu

Neural networks produced by standard training are known to suffer from poor accuracy on rare subgroups despite achieving high accuracy on average, due to the correlations between certain spurious features and labels. Previous approaches based on worst-group loss minimization (\textit{e.g.} Group-DRO) are effective in improving worse-group accuracy but require expensive group annotations for all the training samples. In this paper, we focus on the more challenging and realistic setting where group annotations are only available on a small validation set or are not available at all. We propose \bam, a novel two-stage training algorithm: in the first stage, the model is trained using a \emph{bias amplification} scheme via introducing a learnable \emph{auxiliary variable} for each training sample together with the adoption of squared loss; in the second stage, we upweight the samples that the bias-amplified model misclassifies, and then continue training the same model on the reweighted dataset. Empirically, \bam leads to consistent improvement over its counterparts in worst-group accuracy, resulting in state-of-the-art performance in spurious correlation benchmarks in computer vision and natural language processing. Moreover, we find a simple stopping criterion that completely removes the need for group annotations, with little or no loss in worst-group accuracy.

****

Actionable Recourse Guided by User Preference

Jayanth Yetukuri,Ian Hardy,Yang Liu

The growing popularity of machine learning models has led to their increased application in domains directly impacting human lives. In critical fields such as healthcare, banking, and criminal justice, tools that ensure trust and transparency are vital for the responsible adoption of these models. One such tool is \emph{actionable recourse} (AR) for negatively impacted users. AR describes recommendations of cost-efficient changes to a user's \emph{actionable} features to help them obtain favorable outcomes. Existing approaches for providing recourse optimize for properties such as proximity, sparsity, validity, and distance-based costs. However, an often-overlooked but crucial requirement for actionability is a consideration of \emph{User Preference} to guide the recourse generation process. Moreover, existing works considering a user's preferences require users to precisely specify their costs for taking actions. This requirement raises questions about the practicality of the corresponding solutions due to the high cognitive loads imposed. In this work, we attempt to capture user preferences via soft constraints in three simple forms: \textit{i) scoring continuous features, ii) bounding feature values} and \textit{iii) ranking categorical features}. We propose an optimization framework that is sensitive to {user preference} and a gradient-based approach to identify \emph{User Preferred Actionable Recourse (UP-AR)}.

We empirically demonstrate the proposed approach's superiority in adhering to user preference while maintaining competitive performance in traditional metrics with extensive experiments.

****

Large Learning Rate Matters for Non-Convex Optimization

Amirkeivan Mohtashami,Martin Jaggi,Sebastian U Stich

When training neural networks, it has been widely observed that a large step siz

e is essential in stochastic gradient descent (SGD) for obtaining superior models. However, the effect of large step sizes on the success of SGD is not well understood theoretically.
Several previous works have attributed this success to the stochastic noise present in SGD. However, we show through a novel set of experiments that the stochastic noise is not sufficient to explain good non-convex training, and that instead the effect of a large learning rate itself is essential for obtaining best performance.
We demonstrate the same effects also in the noise-less case, i.e. for full-batch GD. We formally prove that GD with large step size---on certain non-convex function classes---follows a different trajectory than GD with a small step size, which can lead to convergence to a global minimum instead of a local one.
Finally, we also demonstrate the difference in trajectories for small and large learning rates for real neural networks, again observing that large learning rates allow escaping from a local minimum, confirming this behavior is indeed relevant in practice.
****

A Deep Learning Framework for Musical Acoustics Simulations

Jiafeng Chen,George Hagopian,K■vanç Tatar,Victor Zappi

The acoustic modeling of musical instruments is a heavy computational process, often bound to the solution of complex systems of partial differential equations (PDEs). Numerical models can achieve a high level of accuracy, but they may take up to several hours to complete a full simulation, especially in the case of intricate musical mechanisms. The application of deep learning, and in particular of neural operators that learn mappings between function spaces, has the potential to revolutionize how acoustics PDEs are solved and noticeably speed up musical simulations. However, such operators require large datasets, capable of exemplifying the relationship between input parameters (excitation) and output solutions (acoustic wave propagation) per each target musical instrument/configuration. With this work, we present an open-access, open-source framework designed for the generation of numerical musical acoustics datasets and for the training/benchmarking of acoustics neural operators. We first describe the overall structure of the framework and the proposed data generation workflow. Then, we detail the first numerical models that were ported to the framework. Finally, we conclude by sharing some preliminary results obtained by means of training a state-of-the-art neural operator with a dataset generated via the framework. This work is a first step towards the gathering of a research community that focuses on deep learning applied to musical acoustics, and shares workflows and benchmarking tools.
****

Domain Generalization via Heckman-type Selection Models

Hyungu Kahng,Hyungrok Do,Judy Zhong

The domain generalization (DG) setup considers the problem where models are trained on data sampled from multiple domains and evaluated on test domains unseen during training. In this paper, we formulate DG as a sample selection problem where each domain is sampled from a common underlying population through non-random sampling probabilities that correlate with both the features and the outcome. Under this setting, the fundamental iid assumption of the empirical risk minimization (ERM) is violated, so it often performs worse on test domains whose non-random sampling probabilities differ from the domains in the training dataset. We propose a Selection-Guided DG (SGDG) framework to learn the selection probability of each domain and the joint distribution of the outcome and domain selection variables. The proposed SGDG is domain generalizable as it intends to minimize the risk under the population distribution. We theoretically proved that, under certain regular conditions, SGDG can achieve smaller risk than ERM. Furthermore, we present a class of parametric SGDG (HeckmanDG) estimators applicable to continuous, binary, and multinomial outcomes. We also demonstrated its efficacy empirically through simulations and experiments on a set of benchmark datasets comparing with other well-known DG methods.
****

Moving Forward by Moving Backward: Embedding Action Impact over Action Semantics

Kuo-Hao Zeng,Luca Weihs,Roozbeh Mottaghi,Ali Farhadi

A common assumption when training embodied agents is that the impact of taking an action is stable; for instance, executing the ``move ahead'' action will always move the agent forward by a fixed distance, perhaps with some small amount of actuator-induced noise. This assumption is limiting; an agent may encounter settings that dramatically alter the impact of actions: a move ahead action on a wet floor may send the agent twice as far as it expects and using the same action with a broken wheel might transform the expected translation into a rotation. Instead of relying that the impact of an action stably reflects its pre-defined semantic meaning, we propose to model the impact of actions on-the-fly using latent embeddings. By combining these latent action embeddings with a novel, transformer-based, policy head, we design an Action Adaptive Policy (AAP). We evaluate our AAP on two challenging visual navigation tasks in the AI2-THOR and Habitat environments and show that our AAP is highly performant even when faced, at inference-time, with missing actions and, previously unseen, perturbed action spaces. Moreover, we observe significant improvement in robustness against these actions when evaluating in real-world scenarios.

****

Guiding Safe Exploration with Weakest Preconditions

Greg Anderson,Swarat Chaudhuri,Isil Dillig

In reinforcement learning for safety-critical settings, it is often desirable for the agent to obey safety constraints at all points in time, including during training. We present a novel neurosymbolic approach called SPICE to solve this safe exploration problem. SPICE uses an online shielding layer based on symbolic weakest preconditions to achieve a more precise safety analysis than existing tools without unduly impacting the training process. We evaluate the approach on a suite of continuous control benchmarks and show that it can achieve comparable performance to existing safe learning techniques while incurring fewer safety violations. Additionally, we present theoretical results showing that SPICE converges to the optimal safe policy under reasonable assumptions.

****

MetaMD: Principled Optimiser Meta-Learning for Deep Learning

Boyan Gao,Henry Gouk,Jan Stuehmer,massimiliano pontil,Timothy Hospedales

Optimiser design influences learning speed and generalisation in training machine learning models. Several studies have attempted to learn more effective gradient-descent optimisers via solving a bi-level optimisation problem where generalisation error is minimised with respect to optimiser parameters. However, most existing neural network oriented optimiser learning methods are intuitively motivated, without clear theoretical support, and focus on learning implicit biases that improve generalisation, rather than speed of convergence. We take a different perspective starting from mirror descent rather than gradient descent, and meta-learning the corresponding Bregman divergence. Within this paradigm, we formalise a novel meta-learning objective of optimising the rate of convergence. The resulting framework, termed Meta Mirror Descent (MetaMD), learns to accelerate optimisation speed. Unlike many meta-learned neural network optimisers, it also supports convergence guarantees and uniquely does so without requiring validation data. We empirically evaluate our framework on a variety of tasks and architectures in terms of convergence rate and generalisation error and demonstrate strong performance.

****

A Sample Based Method for Understanding The Decisions of Neural Networks Semantically

Ohi Dibua,Jonathan Mbuya,Mackenzie Austin,Kushal Kafle

Interpretability in deep learning is one of the largest obstacles to its more widespread adoption in critical applications. A variety of methods have been introduced to understand and explain decisions made by Deep Models. A class of these methods highlights which features are most influential to model predictions. These methods have some key weaknesses. First, most of these methods are applicable only to the atomic elements that make up raw inputs to the model (e.g., pixels or words). Second, these methods generally don't distinguish between the importa

nce of features individually versus due to interactions with other features. As a result, it is difficult to explore high level questions about how models use features. We tackle these issues by proposing Sample-Based Semantic Analysis (SBSA). We use Sobol sensitivity analysis as our sample-based method. Sobol-SBSA allows us to quantify the importance of semantic combinations of raw inputs and highlight the extent to which these features are important individually as opposed to due to interactions with other features. We demonstrate the ability of Sobol-SBSA to answer a richer class of questions about the behavior of Deep Learning models by exploring how CNN models from AlexNet to DenseNet use regions when classifying images. We present two key findings. 1) The architectural improvements from AlexNet to DenseNet manifested themselves in CNN models utilizing greater levels of region interactions for predictions. 2) Adversarially robust CNNs resist exploiting spurious correlations in ImageNet data by forcing these architectures to rely less on region-to-region interaction. Our proposed method is generalizable to a wide variety of network and input types and can help provide greater clarity about model decisions.
****

Deep Biological Pathway Informed Pathology-Genomic Multimodal Survival Prediction

Lin Qiu,Aminollah Khormali,Kai Liu

The integration of multi-modal data, such as pathological images and genomic data, is essential for understanding cancer heterogeneity and complexity for personalized treatments, as well as for enhancing survival predictions. Despite the progress made in integrating pathology and genomic data, most existing methods cannot mine the complex inter-modality relations thoroughly. Additionally, identifying explainable features from these models that govern preclinical discovery and clinical prediction is crucial for cancer diagnosis, prognosis, and therapeutic response studies. We propose PONET- a novel biological pathway informed pathology-genomic deep model that integrates pathological images and genomic data not only to improve survival prediction but also to identify genes and pathways that cause different survival rates in patients. Empirical results on six of The Cancer Genome Atlas (TCGA) datasets show that our proposed method achieves superior predictive performance and reveals meaningful biological interpretations. The proposed method establishes insight on how to train biological informed deep networks on multimodal biomedical data which will have general applicability for understanding diseases and predicting response and resistance to treatment.
****

A CMDP-within-online framework for Meta-Safe Reinforcement Learning

Vanshaj Khattar,Yuhao Ding,Bilgehan Sel,Javad Lavaei,Ming Jin

Meta-reinforcement learning has widely been used as a learning-to-learn framework to solve unseen tasks with limited experience. However, the aspect of constraint violations has not been adequately addressed in the existing works, making their application restricted in real-world settings. In this paper, we study the problem of meta-safe reinforcement learning (meta-SRL) through the CMDP-within-online framework. We obtain task-averaged regret guarantees for the reward maximization (optimality gap) and constraint violations using gradient-based meta-learning and show that the task-averaged optimality gap and constraint satisfaction improve with task-similarity in the static environment, or task-relatedness in the changing environment. Several technical challenges arise when making this framework practical while still having strong theoretical guarantees. To address these challenges, we propose a meta-algorithm that performs inexact online learning on the upper bounds of intra-task optimality gap and constraint violations estimated by off-policy stationary distribution corrections. Furthermore, we enable the learning rates to be adapted for every task and extend our approach to settings with the dynamically changing task environments. Finally, experiments are conducted to demonstrate the effectiveness of our approach. The proposed theoretical framework is the first to handle the nonconvexity and stochastic nature of within-task CMDPs, while exploiting inter-task dependency for multi-task safe learning.

****

Active Sampling for Node Attribute Completion on Graphs

Benyuan Liu,Xu Chen,Yanfeng Wang,Ya Zhang,Zhi Cao,Ivor Tsang

Node attribute is one kind of crucial information on graphs, but real-world graphs usually face attribute-missing problem where attributes of partial nodes are missing and attributes of the other nodes are available. It is meaningful to restore the missing attributes so as to benefit downstream graph learning tasks. Popular GNN is not designed for this node attribute completion issue and is not capable of solving it. Recent proposed Structure-attribute Transformer (SAT) framework decouples the input of graph structures and node attributes by a distribution matching technique, and can work on it properly. However, SAT leverages nodes with observed attributes in an equally-treated way and neglects the different contributions of different nodes in learning. In this paper, we propose a novel active sampling algorithm (ATS) to more efficiently utilize the nodes with observed attributes and better restore the missing node attributes. Specifically, ATS contains two metrics that measure the representativeness and uncertainty of each node's information by considering the graph structures, representation similarity and learning bias. Then, these two metrics are linearly combined by a Beta distribution controlled weighting scheme to finally determine which nodes are selected into the train set in the next optimization step. This ATS algorithm can be combined with SAT framework together, and is learned in an iterative manner. Through extensive experiments on 4 public benchmark datasets and two downstream tasks, we show the superiority of ATS in node attribute completion.

****

Effects of Graph Convolutions in Multi-layer Networks

Aseem Baranwal,Kimon Fountoulakis,Aukosh Jagannath

Graph Convolutional Networks (GCNs) are one of the most popular architectures that are used to solve classification problems accompanied by graphical information. We present a rigorous theoretical understanding of the effects of graph convolutions in multi-layer networks. We study these effects through the node classification problem of a non-linearly separable Gaussian mixture model coupled with a stochastic block model. First, we show that a single graph convolution expands the regime of the distance between the means where multi-layer networks can classify the data by a factor of at least $1/\sqrt[4]{\rm deg}$, where ${\rm deg}$ denotes the expected degree of a node. Second, we show that with a slightly stronger graph density, two graph convolutions improve this factor to at least $1/\sqrt[4]{n}$, where n is the number of nodes in the graph. Finally, we provide both theoretical and empirical insights into the performance of graph convolutions placed in different combinations among the layers of a neural network, concluding that the performance is mutually similar for all combinations of the placement. We present extensive experiments on both synthetic and real-world data that illustrate our results.

****

SimPer: Simple Self-Supervised Learning of Periodic Targets

Yuzhe Yang,Xin Liu,Jiang Wu,Silviu Borac,Dina Katabi,Ming-Zher Poh,Daniel McDuff

From human physiology to environmental evolution, important processes in nature often exhibit meaningful and strong periodic or quasi-periodic changes. Due to their inherent label scarcity, learning useful representations for periodic tasks with limited or no supervision is of great benefit. Yet, existing self-supervised learning (SSL) methods overlook the intrinsic periodicity in data, and fail to learn representations that capture periodic or frequency attributes. In this paper, we present SimPer, a simple contrastive SSL regime for learning periodic information in data. To exploit the periodic inductive bias, SimPer introduces customized augmentations, feature similarity measures, and a generalized contrastive loss for learning efficient and robust periodic representations. Extensive experiments on common real-world tasks in human behavior analysis, environmental sensing, and healthcare domains verify the superior performance of SimPer compared to state-of-the-art SSL methods, highlighting its intriguing properties including better data efficiency, robustness to spurious correlations, and generalization to distribution shifts.

****

Explaining Patterns in Data with Language Models via Interpretable Autoprompting

Chandan Singh,John Xavier Morris,Jyoti Aneja,Alexander M Rush,Jianfeng Gao

Large language models (LLMs) have displayed an impressive ability to harness natural language to perform complex tasks. In this work, we explore whether we can leverage this learned ability to find and explain patterns in data. Specifically, given a pre-trained LLM and data examples, we introduce interpretable autoprompting (iPrompt), an algorithm that generates a natural-language string explaining the data. iPrompt iteratively alternates between generating explanations with an LLM and reranking them based on their performance when used as a prompt. Experiments on a wide range of datasets, from synthetic mathematics to natural-language understanding, show that iPrompt can yield meaningful insights by accurately finding groundtruth dataset descriptions. Moreover, the prompts produced by iPrompt are simultaneously human-interpretable and highly effective for generalization: on real-world sentiment classification datasets, iPrompt produces prompts that match or even improve upon human-written prompts for GPT-3. Finally, experiments with an fMRI dataset show the potential for iPrompt to aid in scientific discovery.

****

Lipschitz regularized gradient flows and latent generative particles

Hyemin Gu,Panagiota Birmpa,Yannis Pantazis,Markos Katsoulakis,Luc Rey-Bellet

Lipschitz regularized f-divergences are constructed by imposing a bound on the Lipschitz constant of the discriminator in the variational representation. These divergences interpolate between the Wasserstein metric and f-divergences and provide a flexible family of loss functions for non-absolutely continuous (e.g. empirical) distributions, possibly with heavy tails. We first construct Lipschitz regularized gradient flows on the space of probability measures based on these divergences. Examples of such gradient flows are Lipschitz regularized Fokker-Planck and porous medium partial differential equations (PDEs) for the Kullback-Leibler and α-divergences, respectively. The regularization corresponds to imposing a Courant–Friedrichs–Lewy numerical stability condition on the PDEs. For empirical measures, the Lipschitz regularization on gradient flows induces a numerically stable transporter/discriminator particle algorithm, where the generative particles are transported along the gradient of the discriminator. The gradient structure leads to a regularized Fisher information which is the total kinetic energy of the particles and can be used to track the convergence of the algorithm. The Lipschitz regularized discriminator can be implemented via neural network spectral normalization and the particle algorithm generates approximate samples from possibly high-dimensional distributions known only from data. Notably, our particle algorithm can generate synthetic data even in small sample size regimes. A new data processing inequality for the regularized divergence allows us to combine our particle algorithm with representation learning, e.g. autoencoder architectures. The resulting particle algorithm in latent space yields markedly improved generative properties in terms of efficiency and quality of the synthetic samples. From a statistical mechanics perspective the encoding can be interpreted dynamically as learning a better mobility for the generative particles.

****

Post-hoc Concept Bottleneck Models

Mert Yuksekgonul,Maggie Wang,James Zou

Concept Bottleneck Models (CBMs) map the inputs onto a set of interpretable concepts (``the bottleneck'') and use the concepts to make predictions. A concept bottleneck enhances interpretability since it can be investigated to understand what concepts the model "sees" in an input and which of these concepts are deemed important. However, CBMs are restrictive in practice as they require dense concept annotations in the training data to learn the bottleneck. Moreover, CBMs often do not match the accuracy of an unrestricted neural network, reducing the incentive to deploy them in practice. In this work, we address these limitations of CBMs by introducing Post-hoc Concept Bottleneck models (PCBMs). We show that we can turn any neural network into a PCBM without sacrificing model performance while still retaining the interpretability benefits. When concept annotations are not available on the training data, we show that PCBM can transfer concepts from other datasets or from natural language descriptions of concepts via multimodal models. A key benefit of PCBM is that it enables users to quickly debug and update the model to reduce spurious correlations and improve generalization to new distributions. PCBM allows for global model edits, which can be more efficient than previous works on local interventions that fix a specific prediction. Through a model-editing user study, we show that editing PCBMs via concept-level feedback can provide significant performance gains without using data from the target domain or model retraining.

****

Undersampling is a Minimax Optimal Robustness Intervention in Nonparametric Classification

Niladri Shekhar Chatterji,Saminul Haque,Tatsunori Hashimoto

While a broad range of techniques have been proposed to tackle distribution shift, the simple baseline of training on an undersampled balanced dataset often achieves close to state-of-the-art-accuracy across several popular benchmarks. This is rather surprising, since undersampling algorithms discard excess majority group data. To understand this phenomenon, we ask if learning is fundamentally constrained by a lack of minority group samples. We prove that this is indeed the case in the setting of nonparametric binary classification. Our results show that in the worst case, an algorithm cannot outperform undersampling unless there is a high degree of overlap between the train and test distributions (which is unlikely to be the case in real-world datasets), or if the algorithm leverages additional structure about the distribution shift. In particular, in the case of label shift we show that there is always an undersampling algorithm that is minimax optimal. In the case of group-covariate shift we show that there is an undersampling algorithm that is minimax optimal when the overlap between the group distributions is small. We also perform an experimental case study on a label shift dataset and find that in line with our theory, the test accuracy of robust neural network classifiers is constrained by the number of minority samples.

****

Emb-GAM: an Interpretable and Efficient Predictor using Pre-trained Language Models

Chandan Singh,Jianfeng Gao

Deep learning models have achieved impressive prediction performance but often sacrifice interpretability, a critical consideration in high-stakes domains such as healthcare or policymaking. In contrast, generalized additive models (GAMs) can maintain interpretability, but often suffer from poor prediction performance due to their inability to effectively capture feature interactions. In this work, we aim to bridge this gap by using pre-trained large-language models to extract embeddings for each input before learning a linear model in the embedding space. The final model (which we call Emb-GAM) is a transparent, linear function of its input features and feature interactions.

Leveraging the language model allows Emb-GAM to learn far fewer linear coefficients, model larger interactions, and generalize well to novel inputs (e.g. unseen ngrams in text). Across a variety of natural-language-processing datasets, Emb-GAM achieves strong prediction performance without sacrificing interpretability. All code for using Emb-GAM and reproducing our results is made available on github.

****

When Source-Free Domain Adaptation Meets Learning with Noisy Labels

Li Yi,Gezheng Xu,Pengcheng Xu,Jiaqi Li,Ruizhi Pu,Charles Ling,Ian McLeod,Boyu Wang

Recent state-of-the-art source-free domain adaptation (SFDA) methods have focused on learning meaningful cluster structures in the feature space, which have succeeded in adapting the knowledge from source domain to unlabeled target domain without accessing the private source data. However, existing methods rely on the pseudo-labels generated by source models that can be noisy due to domain shift.

In this paper, we study SFDA from the perspective of learning with label noise (LLN). Unlike the label noise in the conventional LLN scenario, we prove that the label noise in SFDA follows a different distribution assumption. We also prove that such a difference makes existing LLN methods that rely on their distribution assumptions unable to address the label noise in SFDA. Empirical evidence suggests that only marginal improvements are achieved when applying the existing LLN methods to solve the SFDA problem. On the other hand, although there exists a fundamental difference between the label noise in the two scenarios, we demonstrate theoretically that the early-time training phenomenon (ETP), which has been previously observed in conventional label noise settings, can also be observed in the SFDA problem. Extensive experiments demonstrate significant improvements to existing SFDA algorithms by leveraging ETP to address the label noise in SFDA.
****
Is a Caption Worth a Thousand Images? A Study on Representation Learning
Shibani Santurkar,Yann Dubois,Rohan Taori,Percy Liang,Tatsunori Hashimoto
The development of CLIP [Radford et al., 2021] has sparked a debate on whether adding language supervision can yield vision models with more transferable representations than traditional image-only methods. Our work studies this question through a carefully controlled comparison of two approaches, in terms of their ability to learn representations that generalize to downstream classification tasks. We find that when the pre-training data meets certain criteria---it is sufficiently large and contains descriptive captions with low variability----image-only methods do not match CLIP's performance even when they are trained with more image data. However, contrary to what one might expect, there are practical settings in which these criteria are not met, wherein added supervision through captions is actually detrimental.
Motivated by our findings, we devise simple data and algorithmic interventions to improve the transfer performance of CLIP-style models.
****
Parameter-Efficient Fine-Tuning Design Spaces
Jiaao Chen,Aston Zhang,Xingjian Shi,Mu Li,Alex Smola,Diyi Yang
Parameter-efficient fine-tuning aims to achieve comparable performances of fine-tuning with much fewer trainable parameters. Recently, various tuning strategies (e.g., Adapters, Prefix Tuning, BitFit, and LoRA) have been proposed. However, their designs are hand-crafted separately, and it remains unclear whether certain design patterns exist for parameter-efficient fine-tuning. Thus, we present a parameter-efficient fine-tuning design paradigm and discover design patterns that are applicable to different experimental settings. Instead of focusing on designing another individual tuning strategy, we introduce parameter-efficient fine-tuning design spaces that parameterize tuning structures and tuning strategies. Specifically, any design space is characterized by four components: layer grouping, trainable parameter allocation, tunable groups, and strategy assignment. Our comprehensive empirical study leads to the discovery of design patterns: (i) grouping layers in a spindle pattern, (ii) uniformly allocating the number of trainable parameters to layers, (ii) tuning all the groups, and (iv) tuning different groups with proper strategies. Our discovered design patterns result in new parameter-efficient fine-tuning methods. Experiments show that these methods consistently outperform investigated parameter-efficient fine-tuning strategies across different backbone models and different tasks in natural language processing.
****
Concept Gradient: Concept-based Interpretation Without Linear Assumption
Andrew Bai,Chih-Kuan Yeh,Neil Y.C. Lin,Pradeep Kumar Ravikumar,Cho-Jui Hsieh
Concept-based interpretations of black-box models are often more intuitive for humans to understand. The most widely adopted approach for concept-based, gradient interpretation is Concept Activation Vector (CAV). CAV relies on learning a linear relation between some latent representation of a given model and concepts. The premise of meaningful concepts lying in a linear subspace of model layers is usually implicitly assumed but does not hold true in general. In this work we proposed Concept Gradient (CG), which extends concept-based, gradient interpretation methods to non-linear concept functions. We showed that for a general (poten

tially non-linear) concept, we can mathematically measure how a small change of concept affects the model’s prediction, which is an extension of gradient-based interpretation to the concept space. We demonstrated empirically that CG outperforms CAV in attributing concept importance on real world datasets and performed case study on a medical dataset. The code is available at github.com/jybai/concept-gradients.

****

Constraining Representations Yields Models That Know What They Don't Know

Joao Monteiro,Pau Rodriguez,Pierre-Andre Noel,Issam H. Laradji,David Vazquez

A well-known failure mode of neural networks is that they may confidently return erroneous predictions. Such unsafe behaviour is particularly frequent when the use case slightly differs from the training context, and/or in the presence of an adversary. This work presents a novel direction to address these issues in a broad, general manner: imposing class-aware constraints on a model's internal activation patterns. Specifically, we assign to each class a unique, fixed, randomly-generated binary vector - hereafter called class code - and train the model so that its cross-depths activation patterns predict the appropriate class code according to the input sample's class. The resulting predictors are dubbed total activation classifiers (TAC), and TACs may either be trained from scratch, or used with negligible cost as a thin add-on on top of a frozen, pre-trained neural network. The distance between a TAC's activation pattern and the closest valid code acts as an additional confidence score, besides the default unTAC'ed prediction head's. In the add-on case, the original neural network's inference head is completely unaffected (so its accuracy remains the same) but we now have the option to use TAC's own confidence and prediction when determining which course of action to take in an hypothetical production workflow. In particular, we show that TAC strictly improves the value derived from models allowed to reject/defer. We provide further empirical evidence that TAC works well on multiple types of architectures and data modalities and that it is at least as good as state-of-the-art alternative confidence scores derived from existing models.

****

Neural Networks Efficiently Learn Low-Dimensional Representations with SGD

Alireza Mousavi-Hosseini,Sejun Park,Manuela Girotti,Ioannis Mitliagkas,Murat A Erdogdu

We study the problem of training a two-layer neural network (NN) of arbitrary width using stochastic gradient descent (SGD) where the input $\boldsymbol{x}\in \mathbb{R}^d$ is Gaussian and the target $y \in \mathbb{R}$ follows a multiple-index model, i.e., $y=g(\langle\boldsymbol{u_1},\boldsymbol{x}\rangle,...,\langle\boldsymbol{u_k},\boldsymbol{x}\rangle)$ with a noisy link function g. We prove that the first-layer weights in the NN converge to the k-dimensional principal subspace spanned by the vectors $\boldsymbol{u_1},...,\boldsymbol{u_k}$ of the true model, when online SGD with weight decay is used for training. This phenomenon has several important consequences when $k \ll d$. First, by employing uniform convergence on this smaller subspace, we establish a generalization error bound of $\mathcal{O}(\sqrt{{kd}/{T}})$ after T iterations of SGD, which is independent of the width of the NN. We further demonstrate that, by recovering the principal direction, SGD-trained ReLU NNs can learn a single-index target of the form $y=f(\langle\boldsymbol{u},\boldsymbol{x}\rangle) + \epsilon$ with a sample complexity linear in d (up to log factors), where f is a monotonic function with at most polynomial growth, and ϵ is the noise. This is in contrast to the known $d^{\Omega(p)}$ samples required to learn any degree p polynomial in the kernel regime, and shows that SGD-trained NNs can outperform the Neural Tangent Kernel at initialization. Finally, we establish compressibility guarantees for NNs using that SGD produces an approximately rank-k first-layer weight matrix.

****

Mixed Federated Learning: Joint Decentralized and Centralized Learning

Sean Augenstein,Andrew Hard,Lin Ning,Karan Singhal,Satyen Kale,Kurt Partridge,Rajiv Mathews

Federated learning (FL) enables learning from decentralized privacy-sensitive da ta, with computations on raw data confined to take place at edge clients. This paper introduces mixed FL, which incorporates an additional loss term calculated at the coordinating server (while maintaining FL’s private data restrictions). For example, additional datacenter data can be leveraged to jointly learn from centralized (datacenter) and decentralized (federated) training data and better match an expected inference data distribution.Mixed FL also enables offloading some intensive computations (e.g., embedding regularization) to the server, greatly reducing communication and client computation load. For these and other mixed FL use cases, we present three algorithms: PARALLEL TRAINING, 1-WAY GRADIENT TRANSFER, and 2-WAY GRADIENT TRANSFER. We perform extensive experiments of the algorithms on three tasks, demonstrating that mixed FL can blend training data to achieve an oracle’s accuracy on an inference distribution, and can reduce communication and computation overhead by more than 90%. Finally, we state convergence bounds for all algorithms, and give intuition on the mixed FL problems best suited to each. The theory confirms our empirical observations of how the algorithms perform under different mixed FL problem settings.

****

OTCOP: Learning optimal transport maps via constraint optimizations

Xiaokai Huo,Hailiang Liu

The approximation power of neural networks makes it an ideal tool to learn optimal transport maps. However, existing methods are mostly based on the Kantorovich duality and require regularization and/or special network structures such as Input Convex Neural Networks (ICNN). In this paper, we propose a direct constraint optimization algorithm for the computation of optimal transport maps based on the Monge formulation. We solve this constraint optimization problem by using three different methods: the penalty method, the augmented Lagrangian method, and the alternating direction method of multipliers method (AMDD). We demonstrate a significant improvement in the accuracy of learned optimal transport maps on benchmarks. Moreover, we show that our methods reduce the regularization effects and accurately learn the target distributions at lower transport cost.

****

An Extensible Multi-modal Multi-task Object Dataset with Materials

Trevor Scott Standley,Ruohan Gao,Dawn Chen,Jiajun Wu,Silvio Savarese

We present EMMa, an Extensible, Multimodal dataset of Amazon product listings that contains rich Material annotations. It contains more than 2.8 million objects, each with image(s), listing text, mass, price, product ratings, and position in Amazon’s product-category taxonomy. We also design a comprehensive taxonomy of 182 physical materials (e.g., Plastic → Thermoplastic → Acrylic). Objects areannotated with one or more materials from this taxonomy. With the numerous attributes available for each object, we develop a Smart Labeling framework to quickly add new binary labels to all objects with very little manual labeling effort, making the dataset extensible. Each object attribute in our dataset can be included in either the model inputs or outputs, leading to combinatorial possibilities in task configurations. For example, we can train a model to predict the object category from the listing text, or the mass and price from the product listing image. EMMa offers a new benchmark for multi-task learning in computer vision and NLP, and allows practitioners to efficiently add new tasks and object attributes at scale.

****

Sampling with Mollified Interaction Energy Descent

Lingxiao Li,qiang liu,Anna Korba,Mikhail Yurochkin,Justin Solomon

Sampling from a target measure whose density is only known up to a normalization constant is a fundamental problem in computational statistics and machine learning. In this paper, we present a new optimization-based method for sampling called mollified interaction energy descent (MIED). MIED minimizes a new class of energies on probability measures called mollified interaction energies (MIEs). These energies rely on mollifier functions---smooth approximations of the Dirac delta originated from PDE theory. We show that as the mollifier approaches the Dirac delta, the MIE converges to the chi-square divergence with respect to the targ et measure and the gradient flow of the MIE agrees with that of the chi-square divergence. Optimizing this energy with proper discretization yields a practical first-order particle-based algorithm for sampling in both unconstrained and constrained domains. We show experimentally that for unconstrained sampling problems our algorithm performs on par with existing particle-based algorithms like SVGD, while for constrained sampling problems our method readily incorporates constrained optimization techniques to handle more flexible constraints with strong performance compared to alternatives.

****

Does Zero-Shot Reinforcement Learning Exist?

Ahmed Touati,Jérémy Rapin,Yann Ollivier

A zero-shot RL agent is an agent that can solve any RL task in a given environment, instantly with no additional planning or learning, after an initial reward-free learning phase. This marks a shift from the reward-centric RL paradigm towards controllable agents that can follow arbitrary instructions in an environment. Current RL agents can solve families of related tasks at best, or require planning anew for each task. Strategies for approximate zero-shot RL have been suggested using successor features (SFs) (Borsa et al., 2018) or forward-backward (FB) representations (Touati & Ollivier, 2021), but testing has been limited.

After clarifying the relationships between these schemes, we introduce improved losses and new SF models, and test the viability of zero-shot RL schemes systematically on tasks from the Unsupervised RL benchmark (Laskin et al., 2021). To disentangle universal representation learning from exploration, we work in an offline setting and repeat the tests on several existing replay buffers.

SFs appear to suffer from the choice of the elementary state features. SFs with Laplacian eigenfunctions do well, while SFs based on auto-encoders, inverse curiosity, transition models, low-rank transition matrix, contrastive learning, or diversity (APS), perform unconsistently. In contrast, FB representations jointly learn the elementary and successor features from a single, principled criterion. They perform best and consistently across the board, reaching 85% of supervised RL performance with a good replay buffer, in a zero-shot manner.

****

Few-Shot Text Classification with Dual Contrastive Consistency Training

Liwen Sun

In this paper, we explore how to utilize pre-trained language model to perform few-shot text classification where only a few annotated examples are given for each class. Since using traditional cross-entropy loss to fine-tune language model under this scenario causes serious overfitting and leads to sub-optimal generalization of model, we adopt supervised contrastive learning on few labeled data and consistency-regularization on vast unlabeled data. Moreover, we propose a novel contrastive consistency to further boost model performance and refine sentence representation. After conducting extensive experiments on four datasets, we demonstrate that our model (FTCC) can outperform state-of-the-art methods and has better robustness.

****

Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability

Alex Damian,Eshaan Nichani,Jason D. Lee

Traditional analyses of gradient descent show that when the largest eigenvalue of the Hessian, also known as the sharpness $S(\theta)$, is bounded by $2/\eta$, training is "stable" and the training loss decreases monotonically. Recent works, however, have observed that this assumption does not hold when training modern neural networks with full batch or large batch gradient descent. Most recently, Cohen at al. (2021) detailed two important phenomena. The first, dubbed \emph{progressive sharpening}, is that the sharpness steadily increases throughout training until it reaches the instability cutoff $2/\eta$. The second, dubbed \emph{edge of stability}, is that the sharpness hovers at $2/\eta$ for the remainder of training while the loss continues decreasing, albeit non-monotonically. We demonstrate that, far from being chaotic, the dynamics of gradient descent at the e dge of stability can be captured by a cubic Taylor expansion: as the iterates diverge in direction of the top eigenvector of the Hessian due to instability, the cubic term in the local Taylor expansion of the loss function causes the curvature to decrease until stability is restored. This property, which we call \emph{self-stabilization}, is a general property of gradient descent and explains its behavior at the edge of stability. A key consequence of self-stabilization is that gradient descent at the edge of stability implicitly follows \emph{projected} gradient descent (PGD) under the constraint $S(\theta) \le 2/\eta$. Our analysis provides precise predictions for the loss, sharpness, and deviation from the PGD trajectory throughout training, which we verify both empirically in a number of standard settings and theoretically under mild conditions. Our analysis uncovers the mechanism for gradient descent's implicit bias towards stability.

****

Conditional Permutation Invariant Flows

Berend Zwartsenberg,Adam Scibior,Matthew Niedoba,Vasileios Lioutas,Justice Sefas,Yunpeng Liu,Setareh Dabiri,Jonathan Wilder Lavington,Trevor Campbell,Frank Wood

We present a novel, conditional generative probabilistic model of set-valued data with a tractable log density. This model is a continuous normalizing flow governed by permutation equivariant dynamics. These dynamics are driven by a learnable per-set-element term and pairwise interactions, both parametrized by deep neural networks. We illustrate the utility of this model via applications including (1) complex traffic scene generation conditioned on visually specified map information, and (2) object bounding box generation conditioned directly on images. We train our model by maximizing the expected likelihood of labeled conditional data under our flow, with the aid of a penalty that ensures the dynamics are smooth and hence efficiently solvable. Our method significantly outperforms non-permutation invariant baselines in terms of log likelihood and domain-specific metrics (offroad, collision, and combined infractions), yielding realistic samples that are difficult to distinguish from real data.

****

TILP: Differentiable Learning of Temporal Logical Rules on Knowledge Graphs

Siheng Xiong,Yuan Yang,Faramarz Fekri,James Clayton Kerce

Compared with static knowledge graphs, temporal knowledge graphs (tKG), which can capture the evolution and change of information over time, are more realistic and general. However, due to the complexity that the notion of time introduces to the learning of the rules, an accurate graph reasoning, e.g., predicting new links between entities, is still a difficult problem. In this paper, we propose TILP, a differentiable framework for temporal logical rules learning. By designing a constrained random walk mechanism and the introduction of temporal operators, we ensure the efficiency of our model. We present temporal features modeling in tKG, e.g., recurrence, temporal order, interval between pair of relations, and duration, and incorporate it into our learning process. We compare TILP with state-of-the-art methods on two benchmark datasets. We show that our proposed framework can improve upon the performance of baseline methods while providing interpretable results. In particular, we consider various scenarios in which training samples are limited, data is biased, and the time range between training and inference are different. In all these cases, TILP works much better than the state-of-the-art methods.

****

Hyperbolic Deep Reinforcement Learning

Edoardo Cetin,Benjamin Paul Chamberlain,Michael M. Bronstein,Jonathan J Hunt

In deep reinforcement learning (RL), useful information about the state is inherently tied to its possible future successors. Consequently, encoding features that capture the hierarchical relationships between states into the model's latent representations is often conducive to recovering effective policies. In this work, we study a new class of deep RL algorithms that promote encoding such relationships by using hyperbolic space to model latent representations. However, we find that a naive application of existing methodology from the hyperbolic deep learning literature leads to fatal instabilities due to the non-stationarity and variance characterizing common gradient estimators in RL. Hence, we design a new

general method that directly addresses such optimization challenges and enables stable end-to-end learning with deep hyperbolic representations. We empirically validate our framework by applying it to popular on-policy and off-policy RL algorithms on the Procgen and Atari 100K benchmarks, attaining near universal performance and generalization benefits. Given its natural fit, we hope this work will inspire future RL research to consider hyperbolic representations as a standard tool.
****

Learning Controllable Adaptive Simulation for Multi-resolution Physics
Tailin Wu,Takashi Maruyama,Qingqing Zhao,Gordon Wetzstein,Jure Leskovec
Simulating the time evolution of physical systems is pivotal in many scientific and engineering problems. An open challenge in simulating such systems is their multi-resolution dynamics: a small fraction of the system is extremely dynamic, and requires very fine-grained resolution, while a majority of the system is changing slowly and can be modeled by coarser spatial scales. Typical learning-based surrogate models use a uniform spatial scale, which needs to resolve to the finest required scale and can waste a huge compute to achieve required accuracy. In this work, we introduce Learning controllable Adaptive simulation for Multi-resolution Physics (LAMP) as the first full deep learning-based surrogate model that jointly learns the evolution model and optimizes appropriate spatial resolutions that devote more compute to the highly dynamic regions. LAMP consists of a Graph Neural Network (GNN) for learning the forward evolution, and a GNN-based actor-critic for learning the policy of spatial refinement and coarsening. We introduce learning techniques that optimizes LAMP with weighted sum of error and computational cost as objective, allowing LAMP to adapt to varying relative importance of error vs. computation tradeoff at inference time. We evaluate our method in a 1D benchmark of nonlinear PDEs and a challenging 2D mesh-based simulation. We demonstrate that our LAMP outperforms state-of-the-art deep learning surrogate models, and can adaptively trade-off computation to improve long-term prediction error: it achieves an average of 33.7% error reduction for 1D nonlinear PDEs, and outperforms MeshGraphNets + classical Adaptive Mesh Refinement (AMR) in 2D mesh-based simulations. Project website with data and code can be found at: http://snap.stanford.edu/lamp.
****

Gated Neural ODEs: Trainability, Expressivity and Interpretability
Timothy Doyeon Kim,Tankut Can,Kamesh Krishnamurthy
Understanding how the dynamics in biological and artificial neural networks implement the computations required for a task is a salient open question in machine learning and neuroscience. In particular, computations requiring complex memory storage and retrieval pose significant challenge for these networks to implement or learn. Recently, a family of models described by neural ordinary differential equations (nODEs) has emerged as powerful dynamical neural network models capable of capturing complex dynamics. Here, we extend nODEs by endowing them with adaptive timescales using gating interactions. We refer to these as gated neural ODEs (gnODEs). Using a task that requires memory of continuous quantities, we demonstrate the inductive bias of the gnODEs to learn (approximate) continuous attractors. We further show how reduced-dimensional gnODEs retain their modeling power while greatly improving interpretability, even allowing explicit visualization of the structure of learned attractors. We introduce a novel measure of expressivity which probes the capacity of a neural network to generate complex trajectories. Using this measure, we explore how the phase-space dimension of the nODEs and the complexity of the function modeling the flow field contribute to expressivity. We see that a more complex function for modeling the flow field allows a lower-dimensional nODE to capture a given target dynamics. Finally, we demonstrate the benefit of gating in nODEs on several real-world tasks.
****

Value-Based Membership Inference Attack on Actor-Critic Reinforcement Learning
Yunhao Yang,ufuk topcu
In actor-critic reinforcement learning (RL), the so-called actor and critic, respectively, compute candidate policies and a value function that evaluates the ca

ndidate policies. Such RL algorithms may be vulnerable to membership inference attacks (MIAs), a privacy attack that infers the data membership, i.e., whether a specific data record belongs to the training dataset. We investigate the vulnerability of value function in actor-critic to MIAs. We develop \textit{CriticAttack}, a new MIA that targets black-box RL agents by examining the correlation between the expected reward and the value function. We empirically show that \textit{CriticAttack} can correctly infer approximately 90\% of the training data membership, i.e., it achieves 90\% attack accuracy. Such accuracy is far beyond the 50\% random guessing accuracy, indicating a severe privacy vulnerability of the value function. To defend against \textit{CriticAttack}, we design a method called \textit{CriticDefense} that inserts uniform noise to the value function. \textit{CriticDefense} can reduce the attack accuracy to 60\% without significantly affecting the agent’s performance.
****

Open-Vocabulary Object Detection upon Frozen Vision and Language Models
Weicheng Kuo,Yin Cui,Xiuye Gu,AJ Piergiovanni,Anelia Angelova

We present F-VLM, a simple open-vocabulary object detection method built uponFrozenVision andLanguageModels. F-VLM simplifies the current multi-stage training pipeline by eliminating the need for knowledge distillation or detection-tailored pretraining. Surprisingly, we observe that a frozen VLM: 1) retains the locality-sensitive features necessary for detection, and 2) is a strong region classifier. We finetune only the detector head and combine the detector and VLM outputs for each region at inference time. F-VLM shows compelling scaling behavior and achieves +6.5 mask AP improvement over the previous state of theart on novel categories of LVIS open-vocabulary detection benchmark. In addition, we demonstrate very competitive results on COCO open-vocabulary detection benchmark and cross-dataset transfer detection, in addition to significant training speed-up and compute savings. Code will be released.

****

Learned Neural Network Representations are Spread Diffusely with Redundancy
Vedant Nanda,Till Speicher,John P Dickerson,Soheil Feizi,Krishna Gummadi,Adrian Weller

Representations learned by pre-training a neural network on a large dataset are increasingly used successfully to perform a variety of downstream tasks. In this work, we take a closer look at how features are encoded in such pre-trained representations. We find that learned representations in a given layer exhibit a degree of diffuse redundancy, ie, any randomly chosen subset of neurons in the layer that is larger than a threshold size shares a large degree of similarity with the full layer and is able to perform similarly as the whole layer on a variety of downstream tasks. For example, a linear probe trained on 20% of randomly picked neurons from a ResNet50 pre-trained on ImageNet1k achieves an accuracy within 5% of a linear probe trained on the full layer of neurons for downstream CIFAR10 classification. We conduct experiments on different neural architectures (including CNNs and Transformers) pre-trained on both ImageNet1k and ImageNet21k and evaluate a variety of downstream tasks taken from the VTAB benchmark. We find that the loss & dataset used during pre-training largely govern the degree of diffuse redundancy and the "critical mass" of neurons needed often depends on the downstream task, suggesting that there is a task-inherent sparsity-performance Pareto frontier. Our findings shed light on the nature of representations learned by pre-trained deep neural networks and suggest that entire layers might not be necessary to perform many downstream tasks. We investigate the potential for exploiting this redundancy to achieve efficient generalization for downstream tasks and also draw caution to certain possible unintended consequences.
****

Neural DAEs: Constrained neural networks
Tue Boesen,Eldad Haber,Uri M. Ascher

In this article we investigate the effect of explicitly adding auxiliary trajectory information to neural networks for dynamical systems. We draw inspiration from the field of differential-algebraic equations and differential equations on m anifolds and implement similar methods in residual neural networks. We discuss constraints through stabilization as well as projection methods, and show when to use which method based on experiments involving simulations of multi-body pendulums and molecular dynamics scenarios. Several of our methods are easy to implement in existing code and have limited impact on performance while giving significant boosts in terms of inference.

****

Revisiting the Assumption of Latent Separability for Backdoor Defenses
Xiangyu Qi,Tinghao Xie,Yiming Li,Saeed Mahloujifar,Prateek Mittal

Recent studies revealed that deep learning is susceptible to backdoor poisoning attacks. An adversary can embed a hidden backdoor into a model to manipulate its predictions by only modifying a few training data, without controlling the training process. Currently, a tangible signature has been widely observed across a diverse set of backdoor poisoning attacks --- models trained on a poisoned dataset tend to learn separable latent representations for poison and clean samples. This latent separation is so pervasive that a family of backdoor defenses directly take it as a default assumption (dubbed latent separability assumption), based on which to identify poison samples via cluster analysis in the latent space. An intriguing question consequently follows: is the latent separation unavoidable for backdoor poisoning attacks? This question is central to understanding whether the assumption of latent separability provides a reliable foundation for defending against backdoor poisoning attacks. In this paper, we design adaptive backdoor poisoning attacks to present counter-examples against this assumption. Our methods include two key components: (1) a set of trigger-planted samples correctly labeled to their semantic classes (other than the target class) that can regularize backdoor learning; (2) asymmetric trigger planting strategies that help to boost attack success rate (ASR) as well as to diversify latent representations of poison samples. Extensive experiments on benchmark datasets verify the effectiveness of our adaptive attacks in bypassing existing latent separation based backdoor defenses. Moreover, our attacks still maintain a high attack success rate with negligible clean accuracy drop. Our studies call for defense designers to take caution when leveraging latent separation as an assumption in their defenses. Our codes are available at https://github.com/Unispac/Circumventing-Backdoor-Defenses.

****

Restricted Strong Convexity of Deep Learning Models with Smooth Activations
Arindam Banerjee,Pedro Cisneros,Libin Zhu,Misha Belkin

We consider the problem of optimization of deep learning models with smooth activation functions. While there exist influential results on the problem from the ``near initialization'' perspective, we shed considerable new light on the problem. In particular, we make two key technical contributions for such models with L layers, m width, and σ_0^2 initialization variance. First, for suitable σ_0^2, we establish a $O(\frac{\text{poly}(L)}{\sqrt{m}})$ upper bound on the spectral norm of the Hessian of such models, considerably sharpening prior results. Second, we introduce a new analysis of optimization based on Restricted Strong Convexity (RSC) which holds as long as the squared norm of the average gradient of predictors is $\Omega(\frac{\text{poly}(L)}{\sqrt{m}})$ for the square loss. We also present results for more general losses. The RSC based analysis does not need the ``near initialization" perspective and guarantees geometric convergence for gradient descent (GD). To the best of our knowledge, ours is the first result on establishing geometric convergence of GD based on RSC for deep learning models, thus becoming an alternative sufficient condition for convergence that does not depend on the widely-used Neural Tangent Kernel (NTK). We share preliminary experimental results supporting our theoretical advances.

****

Koopman Neural Operator Forecaster for Time-series with Temporal Distributional Shifts
Rui Wang,Yihe Dong,Sercan O Arik,Rose Yu

Temporal distributional shifts, with underlying dynamics changing over time, frequently occur in real-world time series and pose a fundamental challenge for dee p neural networks (DNNs). In this paper, we propose a novel deep sequence model based on the Koopman theory for time series forecasting: Koopman Neural Forecaster (KNF) that leverages DNNs to learn the linear Koopman space and the coefficients of chosen measurement functions. KNF imposes appropriate inductive biases for improved robustness against distributional shifts, employing both a global operator to learn shared characteristics and a local operator to capture changing dynamics, as well as a specially-designed feedback loop to continuously update the learnt operators over time for rapidly varying behaviors. We demonstrate that KNF achieves superior performance compared to the alternatives, on multiple time series datasets that are shown to suffer from distribution shifts.

****

Uncertainty-Driven Exploration for Generalization in Reinforcement Learning

Yiding Jiang,J Zico Kolter,Roberta Raileanu

Value-based methods tend to outperform policy optimization methods when trained and tested in single environments; however, they significantly underperform when trained on multiple environments with similar characteristics and tested on new ones from the same distribution. We investigate the potential reasons behind the poor generalization performance of value-based methods and discover that exploration plays a crucial role in these settings. Exploration is helpful not only for finding optimal solutions to the training environments, but also for acquiring knowledge that helps generalization to other unseen environments. We show how to make value-based methods competitive with policy optimization methods in these settings by using uncertainty-driven exploration and distribtutional RL. Our algorithm is the first value-based method to achieve state-of-the-art on both Procgen and Crafter, two challenging benchmarks for generalization in RL.

****

On Convergence of Federated Averaging Langevin Dynamics

Wei Deng,Qian Zhang,Yian Ma,Zhao Song,Guang Lin

We propose a federated averaging Langevin algorithm (FA-LD) for uncertainty quantification and mean predictions with distributed clients. In particular, we generalize beyond normal posterior distributions and consider a general class of models. We develop theoretical guarantees for FA-LD for strongly log-concave distributions with non-i.i.d data and study how the injected noise and the stochastic-gradient noise, the heterogeneity of data, and the varying learning rates affect the convergence. Such an analysis sheds light on the optimal choice of local updates to minimize communication cost. Important to our approach is that the communication efficiency does not deteriorate with the injected noise in the Langevin algorithms. In addition, we examine in our FA-LD algorithm both independent and correlated noise used over different clients. We observe there is a trade-off between the pairs among communication, accuracy, and data privacy. As local devices may become inactive in federated networks, we also show convergence results based on different averaging schemes where only partial device updates are available. In such a case, we discover an additional bias that does not decay to zero.

****

Posthoc Privacy guarantees for neural network queries

Abhishek Singh,Praneeth Vepakomma,Vivek Sharma,Ramesh Raskar

Cloud based machine learning inference is an emerging paradigm where users share their data with a service provider. Due to increased concerns over data privacy, recent works have proposed using Adversarial Representation Learning (ARL) to learn a privacy-preserving encoding of sensitive user data before it is shared with an untrusted service provider. Traditionally, the privacy of these encodings is evaluated empirically as they lack formal guarantees. In this work, we develop a new framework that provides formal privacy guarantees for an arbitrarily trained neural network by linking its local Lipschitz constant with its local sensitivity. To utilize local sensitivity for guaranteeing privacy, we extend the Propose-Test-Release(PTR) framework to make it tractable for neural network based queries. We verify the efficacy of our framework experimentally on real-world datasets and elucidate the role of ARL in improving the privacy-utility tradeoff.

****

MetaGL: Evaluation-Free Selection of Graph Learning Models via Meta-Learning
Namyong Park,Ryan A. Rossi,Nesreen Ahmed,Christos Faloutsos

Given a graph learning task, such as link prediction, on a new graph, how can we select the best method as well as its hyperparameters (collectively called a model) without having to train or evaluate any model on the new graph? Model selection for graph learning has been largely ad hoc. A typical approach has been to apply popular methods to new datasets, but this is often suboptimal. On the other hand, systematically comparing models on the new graph quickly becomes too costly, or even impractical. In this work, we develop the first meta-learning approach for evaluation-free graph learning model selection, called MetaGL, which utilizes the prior performances of existing methods on various benchmark graph datasets to automatically select an effective model for the new graph, without any model training or evaluations. To quantify similarities across a wide variety of graphs, we introduce specialized meta-graph features that capture the structural characteristics of a graph. Then we design G-M network, which represents the relations among graphs and models, and develop a graph-based meta-learner operating on this G-M network, which estimates the relevance of each model to different graphs. Extensive experiments show that using MetaGL to select a model for the new graph greatly outperforms several existing meta-learning techniques tailed for graph learning model selection (up to 47% better), while being extremely fast at test time (~1 sec).

****

FOCUS: Fairness via Agent-Awareness for Federated Learning on Heterogeneous Data
Wenda Chu,Chulin Xie,Boxin Wang,Linyi Li,Lang Yin,Han Zhao,Bo Li

Federated learning (FL) provides an effective collaborative training paradigm, allowing local agents to train a global model jointly without sharing their local data to protect privacy.

On the other hand, due to the heterogeneous nature of local agents, it is challenging to optimize or even define the fairness for agents, which may discourage valuable participation. For instance, the trained global model may sacrifice the performance of a minority user with high-quality data based on loss optimization over all users.

Existing work usually considers accuracy equity as fairness for different users in FL, which is limited especially under the heterogeneous setting, since it is intuitively "unfair" that agents with low-quality data would achieve similar accuracy.

In this work, we aim to address such limitations and propose a formal fairness definition in FL, fairness via agent-awareness (FAA), which takes the heterogeneous data contributions of local agents into account. In addition, we propose a fair FL training algorithm based on agent clustering (FOCUS) to achieve FAA. Theoretically, we prove the convergence and optimality of FOCUS under mild conditions for linear and general convex loss functions with bounded smoothness. We also prove that FOCUS always achieves higher fairness measured by FAA compared with standard FedAvg protocol under both linear and general convex loss functions. Empirically, we evaluate FOCUS on four datasets, including synthetic data, images, and texts under different settings, and we show that FOCUS achieves significantly higher fairness based on FAA while maintaining similar or even higher prediction accuracy compared with FedAvg and other existing fair FL algorithms.

****

Co-Evolution As More Than a Scalable Alternative for Multi-Agent Reinforcement Learning
Patrick Grzywok

In recent years, gradient based multi-agent reinforcement learning is growing in success. One contributing factor is the use of shared parameters for learning policy networks. While this approach scales well with the number of agents during execution it lacks this ambiguity for training as the number of produced samples grows linearly with the number of agents. For a very large number of agents, this could lead to an inefficient use of the circumstantial amount of produced samples. Moreover in single-agent reinforcement learning policy search with evolut

ionary algorithms showed viable success when sampling can be parallelized on a larger scale. The here proposed method does not only consider sampling in concurrent environments but further investigates sampling diverse parameters from the population in co-evolution in joint environments during training. This co-evolutionary policy search has shown to be capable of training a large number of agents. Beyond that, it has been shown to produce competitive results in smaller environments in comparison to gradient descent based methods. This surprising result make evolutionary algorithms a promising candidate for further research in the context of multi-agent reinforcement learning.
****

Adaptive Parametric Prototype Learning for Cross-Domain Few-Shot Classification
Marzi Heidari,Abdullah Alchihabi,Qing En,Yuhong Guo
Cross-domain few-shot classification induces a much more challenging problem than its in-domain counterpart due to the existence of domain shifts between the training and test tasks.
In this paper, we develop a novel Adaptive Parametric Prototype Learning (APPL) method under the meta-learning convention for cross-domain few-shot classification.
Different from existing prototypical few-shot methods that use the averages of support instances to calculate the class prototypes, we propose to learn class prototypes from the concatenated features of the support set in a parametric fashion and meta-learn the model by enforcing prototype-based regularization on the query set.
In addition, we fine-tune the model in the target domain in a transductive manner using a weighted-moving-average self-training approach on the query instances.

We conduct experiments on multiple cross-domain few-shot benchmark datasets.
The empirical results demonstrate that APPL yields superior performance than many state-of-the-art methods.
****

Minimum Description Length Control
Ted Moskovitz,Ta-Chu Kao,Maneesh Sahani,Matthew Botvinick
We propose a novel framework for multitask reinforcement learning based on the minimum description length (MDL) principle. In this approach, which we term MDL-control (MDL-C), the agent learns the common structure among the tasks with which it is faced and then distills it into a simpler representation which facilitates faster convergence and generalization to new tasks. In doing so, MDL-C naturally balances adaptation to each task with epistemic uncertainty about the task distribution. We motivate MDL-C via formal connections between the MDL principle and Bayesian inference, derive theoretical performance guarantees, and demonstrate MDL-C's empirical effectiveness on both discrete and high-dimensional continuous control tasks.
****

RainProof: An Umbrella to Shield Text Generator from Out-Of-Distribution Data
Maxime DARRIN,Pablo Piantanida,Pierre Colombo
As more and more conversational and translation systems are deployed in production, it is essential to implement and develop effective control mechanisms to ensure their proper functioning and security. An essential component to ensure the safe behavior of the system is out-of-distribution (OOD) detection, which aims to detect whether an input sample is statistically far from the training distribution. While OOD detection is a widely covered topic in classification tasks, it has received much less attention in text generation. This paper addresses the problem of OOD detection for machine translation and dialog generation from an operational perspective. Our contribution includes (i) RAINPROOF a Relative informAItioN Projection Out OF distribution detection framework and (ii) a more operational evaluation setting for OOD detection. Surprisingly, we find that OOD detection is not necessarily aligned with task-specific measures. The OOD detector may filter out samples that are well processed by the model and keep samples that are not, leading to weaker performance. Our results show that RAINPROOF breaks this curse and achieve good results in OOD detection while increasing system perfo

rmance.

****

Variance Double-Down: The Small Batch Size Anomaly in Multistep Deep Reinforcement Learning

Johan Samir Obando Ceron,Marc G Bellemare,Pablo Samuel Castro

State of the art results in reinforcement learning suggest that multi-step learning is necessary. However, the increased variance that comes with it makes it difficult to increase the update horizon beyond relatively small numbers. In this paper, we report the counterintuitive finding that decreasing the batch size substantially improves performance across a large swath of deep RL agents. It is well-known that gradient variance decreases with increasing batch sizes, so obtaining improved performance by increasing variance on two fronts is a rather surprising finding. We conduct a broad set of experiments to better understand this variance double-down phenomenon.

****

PerFedMask: Personalized Federated Learning with Optimized Masking Vectors

Mehdi Setayesh,Xiaoxiao Li,Vincent W.S. Wong

Recently, various personalized federated learning (FL) algorithms have been proposed to tackle data heterogeneity. To mitigate device heterogeneity, a common approach is to use masking. In this paper, we first show that using random masking can lead to a bias in the obtained solution of the learning model. To this end, we propose a personalized FL algorithm with optimized masking vectors called PerFedMask. In particular, PerFedMask facilitates each device to obtain its optimized masking vector based on its computational capability before training. Fine-tuning is performed after training. PerFedMask is a generalization of a recently proposed personalized FL algorithm, FedBABU (Oh et al., 2022). PerFedMask can be combined with other FL algorithms including HeteroFL (Diao et al., 2021) and Split-Mix FL (Hong et al., 2022). Results based on CIFAR-10 and CIFAR-100 datasets show that the proposed PerFedMask algorithm provides a higher test accuracy after fine-tuning and lower average number of trainable parameters when compared with six existing state-of-the-art FL algorithms in the literature. The codes are available at https://github.com/MehdiSet/PerFedMask.

****

Variational Latent Branching Model for Off-Policy Evaluation

Qitong Gao,Ge Gao,Min Chi,Miroslav Pajic

Model-based methods have recently shown great potential for off-policy evaluation (OPE); offline trajectories induced by behavioral policies are fitted to transitions of Markov decision processes (MDPs), which are used to rollout simulated trajectories and estimate the performance of policies. Model-based OPE methods face two key challenges. First, as offline trajectories are usually fixed, they tend to cover limited state and action space. Second, the performance of model-based methods can be sensitive to the initialization of their parameters. In this work, we propose the variational latent branching model (VLBM) to learn the transition function of MDPs by formulating the environmental dynamics as a compact latent space, from which the next states and rewards are then sampled. Specifically, VLBM leverages and extends the variational inference framework with the recurrent state alignment (RSA), which is designed to capture as much information underlying the limited training data, by smoothing out the information flow between the variational (encoding) and generative (decoding) part of VLBM. Moreover, we also introduce the branching architecture to improve the model’s robustness against randomly initialized model weights. The effectiveness of the VLBM is evaluated on the deep OPE (DOPE) benchmark, from which the training trajectories are designed to result in varied coverage of the state-action space. We show that the VLBM outperforms existing state-of-the-art OPE methods in general.

****

Discretization Invariant Learning on Neural Fields

Clinton Wang,Polina Golland

While neural fields have emerged as powerful representations of continuous data, there is a need for neural networks that can perform inference on such data without being sensitive to how the field is sampled, a property called discretizati

on invariance. We develop DI-Net, a framework for learning discretization invariant operators on neural fields of any type. Whereas current theoretical analyses of discretization invariant networks are restricted to the limit of infinite samples, our analysis does not require infinite samples and establishes upper bounds on the variation in DI-Net outputs given different finite discretizations. Our framework leads to a family of neural networks driven by numerical integration via quasi-Monte Carlo sampling with discretizations of low discrepancy. DI-Nets enjoy several desirable theoretical properties such as universal approximation of a large class of maps between L^2 functions with gradients that are also discretization invariant. DI-Nets can also be seen as generalizations of many existing network families as they bridge discrete and continuous network classes, such as convolutional neural networks (CNNs) and neural operators respectively. Experimentally, DI-Nets derived from CNNs are demonstrated to classify and segment visual data represented by neural fields under various discretizations, and sometimes even generalize to new types of discretizations at test time.
****

Tuning Frequency Bias in Neural Network Training with Nonuniform Data
Annan Yu,Yunan Yang,Alex Townsend
Small generalization errors of over-parameterized neural networks (NNs) can be partially explained by the frequency biasing phenomenon, where gradient-based algorithms minimize the low-frequency misfit before reducing the high-frequency residuals. Using the Neural Tangent Kernel (NTK), one can provide a theoretically rigorous analysis for training where data are drawn from constant or piecewise-constant probability densities. Since most training data sets are not drawn from such distributions, we use the NTK model and a data-dependent quadrature rule to theoretically quantify the frequency biasing of NN training given fully nonuniform data. By replacing the loss function with a carefully selected Sobolev norm, we can further amplify, dampen, counterbalance, or reverse the intrinsic frequency biasing in NN training.
****

Global Counterfactual Explanations Are Reliable Or Efficient, But Not Both
Dan Ley,Saumitra Mishra,Daniele Magazzeni
Counterfactual explanations have been widely studied in explainability, with a range of application dependent methods emerging in fairness, recourse and model understanding. The major shortcoming associated with these methods, however, is their inability to provide explanations beyond the local or instance-level. While many works touch upon the notion of a global explanation, typically suggesting to aggregate masses of local explanations in the hope of ascertaining global properties, few provide frameworks that are both reliable and computationally tractable. Meanwhile, practitioners are requesting more efficient and interactive explainability tools. We take this opportunity to investigate existing methods, improving the efficiency of Actionable Recourse Summaries (AReS), one of the only known global recourse frameworks, and proposing Global & Efficient Counterfactual Explanations (GLOBE-CE), a novel and flexible framework that tackles the scalability issues associated with current state-of-the-art, particularly on higher dimensional datasets and in the presence of continuous features. Furthermore, we provide a unique mathematical analysis of categorical feature translations, utilising it in our method. Experimental evaluation with real world datasets and user studies verify the speed, reliability and interpretability improvements of our framework.
****

Learning Multimodal Data Augmentation in Feature Space
Zichang Liu,Zhiqiang Tang,Xingjian Shi,Aston Zhang,Mu Li,Anshumali Shrivastava,Andrew Gordon Wilson
The ability to jointly learn from multiple modalities, such as text, audio, and visual data, is a defining feature of intelligent systems. While there have been promising advances in designing neural networks to harness multimodal data, the enormous success of data augmentation currently remains limited to single-modality tasks like image classification. Indeed, it is particularly difficult to augment each modality while preserving the overall semantic structure of the data;

for example, a caption may no longer be a good description of an image after standard augmentations have been applied, such as translation. Moreover, it is challenging to specify reasonable transformations that are not tailored to a particular modality. In this paper, we introduce LeMDA, Learning Multimodal Data Augmentation, an easy-to-use method that automatically learns to jointly augment multimodal data in feature space, with no constraints on the identities of the modalities or the relationship between modalities. We show that LeMDA can (1) profoundly improve the performance of multimodal deep learning architectures, (2) apply to combinations of modalities that have not been previously considered, and (3) achieve state-of-the-art results on a wide range of applications comprised of image, text, and tabular data.

****

Where to Begin? On the Impact of Pre-Training and Initialization in Federated Learning

John Nguyen,Jianyu Wang,Kshitiz Malik,Maziar Sanjabi,Michael Rabbat

An oft-cited challenge of federated learning is the presence of heterogeneity. \emph{Data heterogeneity} refers to the fact that data from different clients may follow very different distributions. \emph{System heterogeneity} refers to client devices having different system capabilities. A considerable number of federated optimization methods address this challenge. In the literature, empirical evaluations usually start federated training from random initialization. However, in many practical applications of federated learning, the server has access to proxy data for the training task that can be used to pre-train a model before starting federated training. Using four standard federated learning benchmark datasets, we empirically study the impact of starting from a pre-trained model in federated learning. Unsurprisingly, starting from a pre-trained model reduces the training time required to reach a target error rate and enables the training of more accurate models (up to 40\%) than is possible when starting from random initialization. Surprisingly, we also find that starting federated learning from a pre-trained initialization reduces the effect of both data and system heterogeneity. We recommend future work proposing and evaluating federated optimization methods to evaluate the performance when starting from random and pre-trained initializations. This study raises several questions for further work on understanding the role of heterogeneity in federated optimization.

****

BigVGAN: A Universal Neural Vocoder with Large-Scale Training

Sang-gil Lee,Wei Ping,Boris Ginsburg,Bryan Catanzaro,Sungroh Yoon

Despite recent progress in generative adversarial network (GAN)-based vocoders, where the model generates raw waveform conditioned on acoustic features, it is challenging to synthesize high-fidelity audio for numerous speakers across various recording environments. In this work, we present BigVGAN, a universal vocoder that generalizes well for various out-of-distribution scenarios without fine-tuning. We introduce periodic activation function and anti-aliased representation into the GAN generator, which brings the desired inductive bias for audio synthesis and significantly improves audio quality. In addition, we train our GAN vocoder at the largest scale up to 112M parameters, which is unprecedented in the literature. We identify and address the failure modes in large-scale GAN training for audio, while maintaining high-fidelity output without over-regularization. Our BigVGAN, trained only on clean speech (LibriTTS), achieves the state-of-the-art performance for various zero-shot (out-of-distribution) conditions, including unseen speakers, languages, recording environments, singing voices, music, and instrumental audio. We release our code and model at: https://github.com/NVIDIA/BigVGAN

****

PaLI: A Jointly-Scaled Multilingual Language-Image Model

Xi Chen,Xiao Wang,Soravit Changpinyo,AJ Piergiovanni,Piotr Padlewski,Daniel Salz,Sebastian Goodman,Adam Grycner,Basil Mustafa,Lucas Beyer,Alexander Kolesnikov,Joan Puigcerver,Nan Ding,Keran Rong,Hassan Akbari,Gaurav Mishra,Linting Xue,Ashish V Thapliyal,James Bradbury,Weicheng Kuo,Mojtaba Seyedhosseini,Chao Jia,Burcu Karagol Ayan,Carlos Riquelme Ruiz,Andreas Peter Steiner,Anelia Angelova,Xiaohua Z hai,Neil Houlsby,Radu Soricut

Effective scaling and a flexible task interface enable large language models to excel at many tasks. We present PaLI, a model that extends this approach to the joint modeling of language and vision. PaLI generates text based on visual and textual inputs, and with this interface performs many vision, language, and multimodal tasks, in many languages. To train PaLI, we make use of large pretrained encoder-decoder language models and Vision Transformers (ViTs). This allows us to capitalize on their existing capabilities and leverage the substantial cost of training them. We find that joint scaling of the vision and language components is important. Since existing Transformers for language are much larger than their vision counterparts, we train a large, 4-billion parameter ViT (ViT-e) to quantify the benefits from even larger-capacity vision models. To train PaLI, we create a large multilingual mix of pretraining tasks, based on a new image-text training set containing 10B images and texts in over 100 languages. PaLI achieves state-of-the-art in multiple vision and language tasks (such as captioning, visual question-answering, scene-text understanding), while retaining a simple, modular, and scalable design.

****

Achieving Sub-linear Regret in Infinite Horizon Average Reward Constrained MDP with Linear Function Approximation

Arnob Ghosh,Xingyu Zhou,Ness Shroff

We study the infinite horizon average reward constrained Markov Decision Process (CMDP). In contrast to existing works on model-based, finite state space, we consider the model-free linear CMDP setup. We first propose a computationally inefficient algorithm and show that $\tilde{\mathcal{O}}(\sqrt{d^3T})$ regret and constraint violation can be achieved, in which T is the number of interactions, and d is the dimension of the feature mapping. We also propose an efficient variant based on the primal-dual adaptation of the LSVI-UCB algorithm and show that $\tilde{\mathcal{O}}((dT)^{3/4})$ regret and constraint violation can be achieved.

This improves the known regret bound of $\tilde{\mathcal{O}}(T^{5/6})$ for the finite state-space model-free constrained RL which was obtained under a stronger assumption compared to ours. We also develop an efficient policy-based algorithm via novel adaptation of the MDP-EXP2 algorithm to our primal-dual set up with $\tilde{\mathcal{O}}(\sqrt{T})$ regret and even zero constraint violation bound under a stronger set of assumptions.

****

Causal Imitation Learning via Inverse Reinforcement Learning

Kangrui Ruan,Junzhe Zhang,Xuan Di,Elias Bareinboim

One of the most common ways children learn when unfamiliar with the environment is by mimicking adults. Imitation learning concerns an imitator learning to behave in an unknown environment from an expert's demonstration; reward signals remain latent to the imitator. This paper studies imitation learning through causal lenses and extends the analysis and tools developed for behavior cloning (Zhang, Kumor, Bareinboim, 2020) to inverse reinforcement learning. First, we propose novel graphical conditions that allow the imitator to learn a policy performing as well as the expert's behavior policy, even when the imitator and the expert's state-action space disagree, and unobserved confounders (UCs) are present. When provided with parametric knowledge about the unknown reward function, such a policy may outperform the expert's. Also, our method is easily extensible and allows one to leverage existing IRL algorithms even when UCs are present, including the multiplicative-weights algorithm (MWAL) (Syed & Schapire, 2008) and the generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016). Finally, we validate our framework by simulations using real-world and synthetic data.

****

Amos: An Adam-style Optimizer with Adaptive Weight Decay towards Model-Oriented Scale

Ran Tian,Ankur P Parikh

We present Amos, a stochastic gradient-based optimizer designed for training deep neural networks. It can be viewed as an Adam optimizer with theoretically supp

orted, adaptive learning-rate decay and weight decay. A key insight behind Amos is that it leverages model-specific information to determine the initial learning-rate and decaying schedules. When used for pre-training BERT variants and T5, Amos consistently converges faster than the state-of-the-art settings of AdamW, achieving better validation loss within <=70% training steps and time, while requiring <=51% memory for slot variables.
****

The Surprising Computational Power of Nondeterministic Stack RNNs
Brian DuSell,David Chiang

Traditional recurrent neural networks (RNNs) have a fixed, finite number of memory cells. In theory (assuming bounded range and precision), this limits their formal language recognition power to regular languages, and in practice, RNNs have been shown to be unable to learn many context-free languages (CFLs). In order to expand the class of languages RNNs recognize, prior work has augmented RNNs with a nondeterministic stack data structure, putting them on par with pushdown automata and increasing their language recognition power to CFLs. Nondeterminism is needed for recognizing all CFLs (not just deterministic CFLs), but in this paper, we show that nondeterminism and the neural controller interact to produce two more unexpected abilities. First, the nondeterministic stack RNN can recognize not only CFLs, but also many non-context-free languages. Second, it can recognize languages with much larger alphabet sizes than one might expect given the size of its stack alphabet. Finally, to increase the information capacity in the stack and allow it to solve more complicated tasks with large alphabet sizes, we propose a new version of the nondeterministic stack that simulates stacks of vectors rather than discrete symbols. We demonstrate perplexity improvements with this new model on the Penn Treebank language modeling benchmark.
****

Critical Initialization of Wide and Deep Neural Networks through Partial Jacobians: General Theory and Applications
Darshil Doshi,Tianyu He,Andrey Gromov

Deep neural networks are notorious for defying theoretical treatment. However, when the number of parameters in each layer tends to infinity the network function is a Gaussian process (GP) and quantitatively predictive description is possible. Gaussian approximation allows to formulate criteria for selecting hyperparameters, such as variances of weights and biases, as well as the learning rate. These criteria rely on the notion of criticality defined for deep neural networks. In this work we describe a new practical way to diagnose criticality. We introduce *partial Jacobians* of a network, defined as derivatives of preactivations in layer l with respect to preactivations in layer $l_0\leq l$. We derive recurrence relations for the norms of partial Jacobians and utilize these relations to analyze criticality of deep fully connected neural networks with LayerNorm and/or residual connections. We derive and implement a simple and cheap numerical test that allows one to select optimal initialization for a broad class of deep neural networks. Using these tools we show quantitatively that proper stacking of the LayerNorm (applied to preactivations) and residual connections leads to an architecture that is critical for any initialization. Finally, we apply our methods to analyze the MLP-Mixer architecture and show that it is everywhere critical.
****

Agnostic Learning of General ReLU Activation Using Gradient Descent
Pranjal Awasthi,Alex Tang,Aravindan Vijayaraghavan

We provide a convergence analysis of gradient descent for the problem of agnostically learning a single ReLU function under Gaussian distributions. Unlike prior work that studies the setting of zero bias, we consider the more challenging scenario when the bias of the ReLU function is non-zero. Our main result establishes that starting from random initialization, in a polynomial number of iterations gradient descent outputs, with high probability, a ReLU function that achieves an error that is within a constant factor of the optimal i.e., it is guaranteed to achieve an error of $O(OPT)$, where OPT is the error of the best ReLU function. This is a significant improvement over existing guarantees for gradient de

scent, which only guarantee error of $O(\sqrt{d \cdot OPT})$ even in the zero-bias case (Frei et al., 2020). We also provide finite sample guarantees, and obtain similar guarantees for a broader class of marginal distributions beyond Gaussians.
****

Parametrizing Product Shape Manifolds by Composite Networks

Josua Sassen,Klaus Hildebrandt,Martin Rumpf,Benedikt Wirth

Parametrizations of data manifolds in shape spaces can be computed using the rich toolbox of Riemannian geometry. This, however, often comes with high computational costs, which raises the question if one can learn an efficient neural network approximation. We show that this is indeed possible for shape spaces with a special product structure, namely those smoothly approximable by a direct sum of low-dimensional manifolds. Our proposed architecture leverages this structure by separately learning approximations for the low-dimensional factors and a subsequent combination. After developing the approach as a general framework, we apply it to a shape space of triangular surfaces. Here, typical examples of data manifolds are given through datasets of articulated models and can be factorized, for example, by a Sparse Principal Geodesic Analysis (SPGA). We demonstrate the effectiveness of our proposed approach with experiments on synthetic data as well as manifolds extracted from data via SPGA.
****

CURE: A Pre-training Framework on Large-scale Patient Data for Treatment Effect Estimation

Ruoqi Liu,Pin-Yu Chen,Ping Zhang

Treatment effect estimation (TEE) refers to the estimation of causal effects, and it aims to compare the difference among treatment strategies on important outcomes. Current machine learning based methods are mainly trained on labeled data with specific treatments or outcomes of interest, which can be sub-optimal if the labeled data are limited. In this paper, we propose a novel transformer-based pre-training and fine-tuning framework called CURE for TEE from observational data. CURE is pre-trained on large-scale unlabeled patient data to learn representative contextual patient representations, and then fine-tuned on labeled patient data for TEE. We design a new sequence encoding for longitudinal (or structured) patient data and we incorporate structure and time into patient embeddings. Evaluated on 4 downstream TEE tasks, CURE outperforms the state-of-the-art methods in terms of an average of 3.8\% and 6.9\% absolute improvement in Area under the ROC Curve (AUC) and Area under the Precision-Recall Curve (AUPR), and 15.7\% absolute improvement in Influence function-based Precision of Estimating Heterogeneous Effects (IF-PEHE). We further demonstrate the data scalability of CURE and verify the results with corresponding randomized clinical trials. Our proposed method provides a new machine learning paradigm for TEE based on observational data.
****

A Probabilistic Approach to Self-Supervised Learning using Cyclical Stochastic Gradient MCMC

Masoumeh Javanbakhat,Christoph Lippert

In this paper we present a practical Bayesian formulation for self-supervised learning method with Cyclical Stochastic Gradient Hamiltonian Monte Carlo (cSGHMC). Within this framework, we place a prior over the parameters of a self-supervised learning model and use cSGHMC to approximate the high dimensional and multimodal posterior distribution over the embeddings. By exploring an expressive posterior over the embeddings, the Bayesian self-supervised learning produces interpretable and diverse representations. Marginalising over these representations results improvement in semi-supervised learning and out-of-distribution detection tasks. We provide experimental results on multiple classification tasks in semi-supervised learning including Cifar10 and Cifar100. Moreover we demonstrate the effectiveness of the proposed method in out-of distribution detection task using SVHN dataset.
****

Tabular Data to Image Generation: Benchmark Data, Approaches, and Evaluation

Alex Tang,Gromit Yeuk-Yin Chan,Ryan A. Rossi,Chang Xiao,Eunyee Koh

In this work, we study the problem of generating a set of images from an arbitrary tabular dataset. The set of generated images provides an intuitive visual summary of the tabular data that can be quickly and easily communicated and understood by the user.

More specifically, we formally introduce this new dataset to image generation task and discuss a few motivating applications including exploratory data analysis and understanding customer segments for creating better marketing campaigns.

We then curate a benchmark dataset for training such models, which we release publicly for others to use and develop new models for other important applications of interest.

Further, we describe a general and flexible framework that serves as a fundamental basis for studying and developing models for this new task of generating images from tabular data.

From the framework, we propose a few different approaches with varying levels of complexity and tradeoffs.

One such approach leverages both numerical and textual data as the input to our image generation pipeline.

The pipeline consists of an image decoder and a conditional auto-regressive sequence generation model which also includes a pre-trained tabular representation in the input layer.

We evaluate the performance of these approaches through several quantitative metrics (FID for image quality and LPIPS scores for image diversity).

****

Learning Hyper Label Model for Programmatic Weak Supervision

Renzhi Wu,Shen-En Chen,Jieyu Zhang,Xu Chu

To reduce the human annotation efforts, the programmatic weak supervision (PWS) paradigm abstracts weak supervision sources as labeling functions (LFs) and involves a label model to aggregate the output of multiple LFs to produce training labels. Most existing label models require a parameter learning step for each dataset. In this work, we present a hyper label model that (once learned) infers the ground-truth labels for each dataset in a single forward pass without dataset-specific parameter learning. The hyper label model approximates an optimal analytical (yet computationally intractable) solution of the ground-truth labels. We train the model on synthetic data generated in the way that ensures the model approximates the analytical optimal solution, and build the model upon Graph Neural Network (GNN) to ensure the model prediction being invariant (or equivariant) to the permutation of LFs (or data points). On 14 real-world datasets, our hyper label model outperforms the best existing methods in both accuracy (by 1.4 points on average) and efficiency (by six times on average). Our code is available at https://github.com/wurenzhi/hyper_label_model

****

SlenderGNN: Accurate, Robust, and Interpretable GNN, and the Reasons for its Success

Jaemin Yoo,Meng-Chieh Lee,Shubhranshu Shekhar,Christos Faloutsos

Can we design a GNN that is accurate and interpretable at the same time? Could it also be robust to handle the case of homophily, heterophily, or even noisy edges without network effects? We propose SlenderGNN that has all desirable properties: (a) accurate, (b) robust, and (c) interpretable. For the reasons of its success, we had to dig deeper: The result is our GNNLIN framework which highlights the fundamental differences among popular GNN models (e.g., feature combination, structural normalization, etc.) and thus reveals the reasons for the success of our SlenderGNN, as well as the reasons for occasional failures of other GNN variants. Thanks to our careful design, SlenderGNN passes all the 'sanity checks' we propose, and it achieves the highest overall accuracy on 9 real-world datasets of both homophily and heterophily graphs, when compared against 10 recent GNN models. Specifically, SlenderGNN exceeds the accuracy of linear GNNs and matches or exceeds the accuracy of nonlinear models with up to 64 times fewer parameters.

****

FedFA: Federated Feature Augmentation
Tianfei Zhou,Ender Konukoglu

Federated learning is a distributed paradigm that allows multiple parties to collaboratively train deep models without exchanging the raw data. However, the data distribution among clients is naturally non-i.i.d., which leads to severe degradation of the learnt model. The primary goal of this paper is to develop a robust federated learning algorithm to address feature shift in clients’ samples, which can be caused by various factors, e.g., acquisition differences in medical imaging. To reach this goal, we propose FedFA to tackle federated learning from a dis- tinct perspective of federated feature augmentation. FedFA is based on a major insight that each client’s data distribution can be characterized by statistics (i.e., mean and standard deviation) of latent features; and it is likely to manipulate these local statistics globally, i.e., based on information in the entire federation, to let clients have a better sense of the underlying distribution and therefore alleviate local data bias. Based on this insight, we propose to augment each local feature statistic probabilistically based on a normal distribution, whose mean is the original statistic and variance quantifies the augmentation scope. Key to our approach is the determination of a meaningful Gaussian variance, which is accomplished by taking into account not only biased data of each individual client, but also underlying feature statistics characterized by all participating clients. We offer both theoretical and empirical justifications to verify the effectiveness of FedFA. Our code is available at https://github.com/tfzhou/FedFA.

■

****

Show and Write: Entity-aware Article Generation with Image Information
Zhongping Zhang,Yiwen Gu,Bryan A. Plummer

Prior work for article generation has primarily focused on generating articles using a human-written prompt to provide topical context and metadata about the article. However, for many applications, such as generating news stories, these articles are also often paired with images and their captions or alt-text, which in turn are based on real-world events and may reference many different named entities that are difficult to be correctly recognized and predicted by language models. To address this shortcoming, this paper introduces an ENtity-aware article Generation method with Image iNformation, ENGIN, to incorporate an article's image information into language models. ENGIN represents articles that can be conditioned on metadata used by prior work and information such as captions and named entities extracted from images. Our key contribution is a novel Entity-aware mechanism to help our model recognize and predict the entity names in articles. We perform experiments on three public datasets, GoodNews, VisualNews, and WikiText. Quantitative results show that our approach improves generated article perplexity by 4-5 points over the base models. Qualitative results demonstrate the text generated by ENGIN is more consistent with embedded article images. We also perform article quality annotation experiments on the generated articles to validate that our model produces higher-quality articles. Finally, we investigate the effect ENGIN has on methods that automatically detect machine-generated articles.

****

Noise$^+$2Noise: Co-taught De-noising Autoencoders for Time-Series Data
Harry Rubin-Falcone,Joyce Lee,Jenna Wiens

We consider the task of learning to recover clean signals given only access to noisy data. Recent work in computer vision has addressed this problem in the context of images using denoising autoencoders (DAEs). However, to date DAEs for learning from noisy data have not been explored in the context of time-series data. DAEs for denoising images often rely on assumptions unlikely to hold in the context of time series, \textit{e.g.}, multiple noisy samples of the same example. Here, we adapt DAEs to cleaning time-series data with noisy samples only. To recover the clean target signal when only given access to noisy target data, we leverage a noise-free auxiliary time-series signal that is related to the target signal. In addition to leveraging the relationship between the target signal an d auxiliary signal, we iteratively filter and learn from clean samples using an approach based on co-teaching. Applied to the task of recovering carbohydrate values for blood glucose management, our approach reduces noise (MSE) in patient-reported carbohydrates from 72g^2 (95\% CI: 54,93) to 18g^2 (13,25), outperforming the best baseline (MSE = 33g^2 (27,43)). We demonstrate strong time-series denoising performance, extending the applicability of DAEs to a previously under-explored setting.

****

Adversarial Representation Learning for Canonical Correlation Analysis
Feng Bao,Steve J. Altschuler,Lani F. Wu

Canonical correlation analysis (CCA) provides a framework to map multimodality data into a maximally correlated latent space. The deep version of CCA has replaced linear maps with deep transformations to enable more flexible correlated data representations; however, this approach requires optimization over all samples for each iteration and poorly scales. Here, we present a deep, adversarial approach to CCA, adCCA, that can be efficiently solved by standard mini-batch training. We reformulate CCA under the constraint that the different modalities are embedded with identical latent distributions, derive a tractable deep CCA target, and use an adversarial framework to efficiently learn the canonical representations. A consequence of the new formation is that adCCA learns maximally correlated representations across multimodalities meanwhile preserves structure within individual modalities. Further, adCCA removes the need for feature transformation and normalization and can be directly applied to diverse modalities and feature encodings. Numerical studies show that the performance of adCCA is robust to data transformations, binary encodings, and corruptions. Together, adCCA provides a scalable approach to align data across modalities without compromising structure within each modality.

****

Neural Implicit Manifold Learning for Topology-Aware Generative Modelling
Brendan Leigh Ross,Gabriel Loaiza-Ganem,Anthony L. Caterini,Jesse C Cresswell

Natural data observed in $\mathbb{R}^n$ is often constrained to an m-dimensional manifold $\mathcal{M}$, where $m < n$. Current probabilistic models represent this manifold by mapping an m-dimensional latent variable through a neural network $f_\theta: \mathbb{R}^m \to \mathbb{R}^n$. Such procedures, which we call pushforward models, incur a straightforward limitation: manifolds cannot in general be represented with a single parameterization, meaning that attempts to do so will incur either computational instability or the inability to learn probability densities within the manifold. To remedy this problem, we propose to model $\mathcal{M}$ as a neural implicit manifold: the set of zeros of a neural network. To learn the data distribution within $\mathcal{M}$, we introduce constrained energy-based models, which use a constrained variant of Langevin dynamics to train and sample within a learned manifold. The resulting model can be manipulated with an arithmetic of manifolds, which allows practitioners to take unions and intersections of model manifolds. In experiments on synthetic and natural data, we show that constrained EBMs can learn manifold-supported distributions with complex topologies more accurately than pushforward models.

****

LT-SNN: Self-Adaptive Spiking Neural Network for Event-based Classification and Object Detection
Ahmed Hasssan,Jian Meng,Jae-sun Seo

Spiking neural networks (SNNs) have received increasing attention due to its high biological plausibility and energy efficiency. The binary spike-based information propagation enables efficient sparse computation with event-based computer vision applications. Prior works investigated direct SNN training algorithm to overcome the non-differentiability of spike generation. However, most of the existing works employ a fixed threshold value for the membrane potential throughout the entire training process, which limits the dynamics of SNNs towards further optimizing the performance. The adaptiveness in the membrane potential threshold and the mismatched mechanism between SNN and biological nervous system remain under-explored in prior works. In this work, we propose LT-SNN, a novel SNN trainin

g algorithm with self-adaptive learnable potential threshold to improve SNN performance. LT-SNN optimizes the layer-wise threshold value throughout SNN training, imitating the self-adaptiveness of the biological nervous system. To stabilize the SNN training even further, we propose separate surrogate gradient path (SGP), a simple-yet-effective method that enables the smooth learning process of SNN training. We validate the proposed LT-SNN algorithm on multiple event-based datasets, including both image classification and object detection tasks. Equipped with high adaptiveness that fully captures the dynamics of SNNs, LT-SNN achieves state-of-the-art performance with compact models. The proposed LT-SNN based classification network surpasses SoTA methods where we achieved 2.71% higher accuracy together with 10.48× smaller model size. Additionally, our LT-SNN-YOLOv2 object detection model demonstrates 0.11 mAP improvement compared to the SoTA SNN-based object detection.
****

Characterizing neural representation of cognitively-inspired deep RL agents during an evidence accumulation task

James Mochizuki-Freeman,Sahaj Singh Maini,Zoran Tiganj

Evidence accumulation is thought to be fundamental for decision-making in humans and other mammals. It has been extensively studied in neuroscience and cognitive science with the goal of explaining how sensory information is sequentially sampled until sufficient evidence has accumulated to favor one decision over others. Neuroscience studies suggest that the hippocampus encodes a low-dimensional ordered representation of evidence through sequential neural activity. Cognitive modelers have proposed a mechanism by which such sequential activity could emerge through the modulation of recurrent weights with a change in the amount of evidence. This gives rise to neurons tuned to a specific magnitude of evidence which resemble neurons recorded in the hippocampus. Here we integrated a cognitive science model inside a deep Reinforcement Learning (RL) agent and trained the agent to perform a simple evidence accumulation task inspired by the behavioral experiments on animals. We compared the agent's performance with the performance of agents equipped with GRUs and RNNs. We found that the agent based on a cognitive model was able to learn much faster and generalize better while having significantly fewer parameters. We also compared the emergent neural activity across agents and found that in some cases, GRU-based agents developed similar neural representations to agents based on a cognitive model. This study illustrates how integrating cognitive models and deep learning systems can lead to brain-like neural representations that can improve learning.
****

Epistemological Bias As a Means for the Automated Detection of Injustices in News Media

Kenya S. Andrews,Lamogha Chiazor

Injustice occurs when someone experiences unfair treatment or their rights are violated. In the context of news media, injustices represent a form of bias through which discriminatory narratives can arise and spread. The automated identification of injustice in text has received little attention, due in part to the fact that underlying stereotypes are rarely explicitly stated and that instances often occur unconsciously due to the pervasive nature of prejudice in society. Here, we leverage the combined use of a fine-tuned BERT-based bias detection model, two stereotype detection models, and a lexicon-based approach to show that epistemological biases (i.e., words, which through their use, presupposes, entails, asserts, hedges, or boosts text to erode or assert a person's capacity as a knower) can assist with the automatic detection of injustice in text.
****

Neural Constraint Inference: Inferring Energy Constraints in Interacting Systems

Armand Comas,Yilun Du,Sandesh Ghimire,Christian Fernandez Lopez,Mario Sznaier,Joshua B. Tenenbaum,Octavia Camps

Systems consisting of interacting agents are prevalent in the world, ranging from dynamical systems in physics to complex biological networks. To build systems which can interact robustly in the real world, it is thus important to be able to infer the precise interactions governing such systems. Existing approaches ty pically discover such interactions by explicitly modeling the feedforward dynamics of the trajectories. In this work, we propose Neural Constraint Inference (NCI) model as an alternative approach to discover such interactions: it discovers a set of relational constraints, represented as energy functions, which when optimized reconstruct the original trajectory. We illustrate how NCI can faithfully predict future trajectory dynamics, achieving more consistent long-rollouts than existing approaches. We show that the constraints discovered by NCI are disentangled and may be intermixed with constraints from other trajectories. Finally, we illustrate how those constraints enable the incorporation of external test-time constraints.

****

Self-supervised Continual Learning based on Batch-mode Novelty Detection

Jingbo Sun,Jiaxin Zhang,Frank Y Liu,Mahantesh M Halappanavar,Deliang Fan,Yu Cao

Continual learning (CL) plays a key role in dynamic systems in order to adapt to new tasks, while preserving previous knowledge. Most existing CL approaches focus on learning new knowledge in a supervised manner, while leaving the data gathering phase to the novelty detection (ND) algorithm. Such presumption limits the practical usage where new data needs to be quickly learned without being labeled. In this paper, we propose a unified approach of CL and ND, in which each new class of the out-of-distribution (ODD) data is first detected and then added to previous knowledge. Our method has three unique features: (1) a unified framework seamlessly tackling both ND and CL problems; (2) a self-supervised method for model adaptation, without the requirement of new data annotation; (3) batch-mode data feeding that maximizes the separation of new knowledge vs. previous learning, which in turn enables high accuracy in continual learning. By learning one class at each step, the new method achieves robust continual learning and consistently outperforms state-of-the-art CL methods in the single-head evaluation on MNIST, CIFAR-10, CIFAR-100 and TinyImageNet datasets.

****

Stable Optimization of Gaussian Likelihoods

Denis Megerle,Fabian Otto,Michael Volpp,Gerhard Neumann

Uncertainty-aware modeling has emerged as a key component in modern machine learning frameworks. The de-facto standard approach adopts heteroscedastic Gaussian distributions and minimizes the negative log-likelihood (NLL) under observed data. However, optimizing this objective turns out to be surprisingly intricate, and the current state-of-the-art reports several instabilities. This work breaks down the optimization problem, initially focusing on non-contextual settings where convergence can be analyzed analytically. We show that (1) in this learning scheme, the eigenvalues of the predictive covariance define stability in learning, and (2) coupling of gradients and predictions build up errors in both mean and covariance if either is poorly approximated. Building on these insights, we propose Trustable, a novel optimizer that overcomes instabilities methodically by combining systematic update restrictions in the form of trust regions with structured, tractable natural gradients. We demonstrate in several challenging experiments that Trustable outperforms current optimizers in regression with neural networks in terms of the NLL, MSE, and further performance metrics. Unlike other optimizers, Trustable yields an improved and more stable fit and can also be applied to multivariate outputs with full covariance matrices.

****

Representing Latent Dimensions Using Compressed Number Lines

Sahaj Singh Maini,James Mochizuki-Freeman,Chirag Shankar Indi,Brandon G Jacques,Per B Sederberg,Marc Howard,Zoran Tiganj

Humans use log-compressed number lines to represent different quantities, including elapsed time, traveled distance, numerosity, sound frequency, etc. Inspired by recent cognitive science and computational neuroscience work, we developed a neural network that learns to construct log-compressed number lines. The network computes a discrete approximation of a real-domain Laplace transform using an RNN with analytically derived weights giving rise to a log-compressed timeline of the past. The network learns to extract latent variables from the input and uses them for global modulation of the recurrent weights turning a timeline into a number line over relevant dimensions. The number line representation greatly simplifies learning on a set of problems that require learning associations in different spaces - problems that humans can typically solve easily. This approach illustrates how combining deep learning with cognitive models can result in systems that learn to represent latent variables in a brain-like manner and exhibit human-like behavior manifested through Weber-Fechner law.

****

Efficient Sequence Packing without Cross-contamination: Accelerating Large Language Models without Impacting Performance

Mario Michael Krell,Matej Kosec,Sergio P. Perez,Andrew William Fitzgibbon

Effective training of today's large language models (LLMs) depends on large batches and long sequences for throughput and accuracy. To handle variable-length sequences on hardware accelerators, it is common practice to introduce padding tokens, so that all sequences in a batch have the same length. We show in this paper that the variation in sequence lengths in common NLP datasets is such that up to 50% of all tokens can be padding. In less common, but not extreme, cases (e.g. GLUE-COLA with sequence length 128), the ratio is up to 89%. Existing methods to address the resulting inefficiency are complicated by the need to avoid "cross-contamination" in self-attention, by a reduction in accuracy when sequence ordering information is lost, or by customized kernel implementations only valid for specific accelerators.

This paper introduces a new formalization of sequence packing in the context of the well-studied bin packing problem, and presents new algorithms based on this formulation which, for example, confer a 2x speedup for phase 2 pretraining in BERT while preserving downstream performance. We show how existing models can be adapted to ensure mathematical equivalence between the original and packed models, meaning that packed models can be trained with existing pre-training and fine-tuning practices.

****

Cortically motivated recurrence enables task extrapolation

Vijay Veerabadran,Yuan Tang,Ritik Raina,Virginia R. de Sa

Feedforward deep neural networks have become the standard class of models in the field of computer vision. Yet, they possess a striking difference relative to their biological counterparts which predominantly perform “recurrent” computations. Why do biological neurons evolve to employ recurrence pervasively? In this paper, we show that a recurrent network is able to flexibly adapt its computational budget during inference and generalize within-task across difficulties. Simultaneously in this study, we contribute a recurrent module we call LocRNN that is designed based on a prior computational model of local recurrent intracortical connections in primates to support such dynamic task extrapolation. LocRNN learns highly accurate solutions to the challenging visual reasoning problems of Mazes and PathFinder that we use here. More importantly, it is able to flexibly use less or more recurrent iterations during inference to zero-shot generalize to less- and more difficult instantiations of each task without requiring extra training data, a potential functional advantage of recurrence that biological visual systems capitalize on. Feedforward networks on the other hand with their fixed computational graphs only partially exhibit this trend, potentially owing to image-level similarities across difficulties. We also posit an intriguing tradeoff between recurrent networks’ representational capacity and their stability in the recurrent state space. Our work encourages further study of the role of recurrence in deep learning models – especially from the context of out-of-distribution generalization & task extrapolation – and their properties of task performance and stability.

****

Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning?

Rui Wen,Zhengyu Zhao,Zhuoran Liu,Michael Backes,Tianhao Wang,Yang Zhang

Indiscriminate data poisoning can decrease the clean test accuracy of a deep learning model by slightly perturbing its training samples.

There is a consensus that such poisons can hardly harm adversarially-trained (AT

) models when the adversarial training budget is no less than the poison budget, i.e., $\epsilon_\mathrm{adv}\geq\epsilon_\mathrm{poi}$. This consensus, however, is challenged in this paper based on our new attack strategy that induces \textit{entangled features} (EntF). The existence of entangled features makes the poisoned data become less useful for training a model, no matter if AT is applied or not. We demonstrate that for attacking a CIFAR-10 AT model under a reasonable setting with $\epsilon_\mathrm{adv}=\epsilon_\mathrm{poi}=8/255$, our EntF yields an accuracy drop of 13.31%, which is $7\times$ better than existing methods and equal to discarding 83% training data. We further show the generalizability of EntF to more challenging settings, e.g., higher AT budgets, partial poisoning, unseen model architectures, and stronger (ensemble or adaptive) defenses. We finally provide new insights into the distinct roles of non-robust vs. robust features in poisoning standard vs. AT models and demonstrate the possibility of using a hybrid attack to poison standard and AT models simultaneously. Our code is available at~\url{https://github.com/WenRuiUSTC/EntF}.

****

Offline Congestion Games: How Feedback Type Affects Data Coverage Requirement

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

This paper investigates when one can efficiently recover an approximate Nash Equilibrium (NE) in offline congestion games. The existing dataset coverage assumption in offline general-sum games inevitably incurs a dependency on the number of actions, which can be exponentially large in congestion games. We consider three different types of feedback with decreasing revealed information. Starting from the facility-level (a.k.a., semi-bandit) feedback, we propose a novel one-unit deviation coverage condition and show a pessimism-type algorithm that can recover an approximate NE. For the agent-level (a.k.a., bandit) feedback setting, interestingly, we show the one-unit deviation coverage condition is not sufficient. On the other hand, we convert the game to multi-agent linear bandits and show that with a generalized data coverage assumption in offline linear bandits, we can efficiently recover the approximate NE. Lastly, we consider a novel type of feedback, the game-level feedback where only the total reward from all agents is revealed. Again, we show the coverage assumption for the agent-level feedback setting is insufficient in the game-level feedback setting, and with a stronger version of the data coverage assumption for linear bandits, we can recover an approximate NE. Together, our results constitute the first study of offline congestion games and imply formal separations between different types of feedback.

****

Learning with Stochastic Orders

Carles Domingo-Enrich,Yair Schiff,Youssef Mroueh

Learning high-dimensional distributions is often done with explicit likelihood modeling or implicit modeling via minimizing integral probability metrics (IPMs). In this paper, we expand this learning paradigm to stochastic orders, namely, the convex or Choquet order between probability measures. Towards this end, exploiting the relation between convex orders and optimal transport, we introduce the Choquet-Toland distance between probability measures, that can be used as a drop-in replacement for IPMs. We also introduce the Variational Dominance Criterion (VDC) to learn probability measures with dominance constraints, that encode the desired stochastic order between the learned measure and a known baseline. We analyze both quantities and show that they suffer from the curse of dimensionality and propose surrogates via input convex maxout networks (ICMNs), that enjoy parametric rates. We provide a min-max framework for learning with stochastic orders and validate it experimentally on synthetic and high-dimensional image generation, with promising results. Finally, our ICMNs class of convex functions and its derived Rademacher Complexity are of independent interest beyond their application in convex orders. Code to reproduce experimental results is available at https://github.com/yair-schiff/stochastic-orders-ICMN.

****

MEDFAIR: Benchmarking Fairness for Medical Imaging

Yongshuo Zong,Yongxin Yang,Timothy Hospedales

A multitude of work has shown that machine learning-based medical diagnosis syst ems can be biased against certain subgroups of people. This has motivated a growing number of bias mitigation algorithms that aim to address fairness issues in machine learning. However, it is difficult to compare their effectiveness in medical imaging for two reasons. First, there is little consensus on the criteria to assess fairness. Second, existing bias mitigation algorithms are developed under different settings, e.g., datasets, model selection strategies, backbones, and fairness metrics, making a direct comparison and evaluation based on existing results impossible. In this work, we introduce MEDFAIR, a framework to benchmark the fairness of machine learning models for medical imaging. MEDFAIR covers eleven algorithms from various categories, ten datasets from different imaging modalities, and three model selection criteria. Through extensive experiments, we find that the under-studied issue of model selection criterion can have a significant impact on fairness outcomes; while in contrast, state-of-the-art bias mitigation algorithms do not significantly improve fairness outcomes over empirical risk minimization (ERM) in both in-distribution and out-of-distribution settings. We evaluate fairness from various perspectives and make recommendations for different medical application scenarios that require different ethical principles. Our framework provides a reproducible and easy-to-use entry point for the development and evaluation of future bias mitigation algorithms in deep learning. Code is available at https://github.com/ys-zong/MEDFAIR.

****

Does Learning from Decentralized Non-IID Unlabeled Data Benefit from Self Supervision?

Lirui Wang,Kaiqing Zhang,Yunzhu Li,Yonglong Tian,Russ Tedrake

The success of machine learning relies heavily on massive amounts of data, which are usually generated and stored across a range of diverse and distributed data sources. Decentralized learning has thus been advocated and widely deployed to make efficient use of distributed datasets, with an extensive focus on supervised learning (SL) problems. Unfortunately, the majority of real-world data are unlabeled and can be highly heterogeneous across sources. In this work, we carefully study decentralized learning with unlabeled data through the lens of self-supervised learning (SSL), specifically contrastive visual representation learning. We study the effectiveness of a range of contrastive learning algorithms under a decentralized learning setting, on relatively large-scale datasets including ImageNet-100, MS-COCO, and a new real-world robotic warehouse dataset. Our experiments show that the decentralized SSL (Dec-SSL) approach is robust to the heterogeneity of decentralized datasets, and learns useful representation for object classification, detection, and segmentation tasks, even when combined with the simple and standard decentralized learning algorithm of Federated Averaging (FedAvg). This robustness makes it possible to significantly reduce communication and to reduce the participation ratio of data sources with only minimal drops in performance. Interestingly, using the same amount of data, the representation learned by Dec-SSL can not only perform on par with that learned by centralized SSL which requires communication and excessive data storage costs, but also sometimes outperform representations extracted from decentralized SL which requires extra knowledge about the data labels. Finally, we provide theoretical insights into understanding why data heterogeneity is less of a concern for Dec-SSL objectives, and introduce feature alignment and clustering techniques to develop a new Dec-SSL algorithm that further improves the performance, in the face of highly non-IID data. Our study presents positive evidence to embrace unlabeled data in decentralized learning, and we hope to provide new insights into whether and why decentralized SSL is effective and/or even advantageous.

****

Polarity is all you need to learn and transfer faster

Qingyang Wang,Michael Alan Powell,Ali Geisa,Eric W Bridgeford,Joshua T Vogelstein

Natural intelligences (NIs) thrive in a dynamic world – they learn quickly, sometimes with only a few samples. In contrast, Artificial intelligence (AI) has achieved supra (-human) level performance in certain AI settings, typically dependent on a prohibitive amount of training samples and computational power. What des ign principle difference between NI and AI could contribute to such a discrepancy? Here, we propose a research avenue based on a simple observation from NIs: post-development, neuronal connections in the brain rarely see polarity switch. Why? Our answer is: to learn and transfer more efficiently. We demonstrate with theory and simulations that if weight polarities are adequately set $\textit{a priori}$, then networks learn with less time and data. We extend such findings onto image classification tasks and demonstrate that polarity, not weight, is a more effective medium for knowledge transfer between networks. We also explicitly illustrate situations in which $\textit{a priori}$ setting the weight polarities is disadvantageous for networks. Our work illustrates the value of weight polarities from the perspective of statistical and computational efficiency for both NI and AI.

****

On the Geometry of Reinforcement Learning in Continuous State and Action Spaces

Saket Tiwari,Omer Gottesman,George Konidaris

Advances in reinforcement learning have led to its successful application in complex tasks with continuous state and action spaces. Despite these advances in practice, most theoretical work pertains to finite state and action spaces. We propose building a theoretical understanding of continuous state and action spaces by employing a geometric lens. Central to our work is the idea that the transition dynamics induce a low dimensional manifold of reachable states embedded in the high-dimensional nominal state space. We prove that, under certain conditions, the dimensionality of this manifold is at most the dimensionality of the action space plus one. This is the first result of its kind, linking the geometry of the state space to the dimensionality of the action space. We empirically corroborate this upper bound for four MuJoCo environments.We further demonstrate the applicability of our result by learning a policy in this low dimensional representation.   To do so we introduce an algorithm that learns a mapping to a low dimensional representation, as a narrow hidden layer of a deep neural network, in tandem with the policy using DDPG. Our experiments show that a policy learnt this way perform on par or better for four MuJoCo control suite tasks.

****

Malign Overfitting: Interpolation and Invariance are Fundamentally at Odds

Yoav Wald,Gal Yona,Uri Shalit,Yair Carmon

Learned classifiers should often possess certain invariance properties meant to encourage fairness, robustness, or out-of-distribution generalization.
However, multiple recent works empirically demonstrate that common invariance-inducing regularizers are ineffective in the over-parameterized regime, in which classifiers perfectly fit (i.e. interpolate) the training data. This suggests that the phenomenon of ``benign overfitting," in which models generalize well despite interpolating, might not favorably extend to settings in which robustness or fairness are desirable.

In this work, we provide a theoretical justification for these observations. We prove that---even in the simplest of settings---any interpolating learning rule (with an arbitrarily small margin) will not satisfy these invariance properties. We then propose and analyze an algorithm that---in the same setting---successfully learns a non-interpolating classifier that is provably invariant. We validate our theoretical observations on simulated data and the Waterbirds dataset.

****

Exploring and Exploiting Decision Boundary Dynamics for Adversarial Robustness

Yuancheng Xu,Yanchao Sun,Micah Goldblum,Tom Goldstein,Furong Huang

The robustness of a deep classifier can be characterized by its margins: the decision boundary's distances to natural data points. However, it is unclear whether existing robust training methods effectively increase the margin for each vulnerable point during training. To understand this, we propose a continuous-time framework for quantifying the relative speed of the decision boundary with respect to each individual point. Through visualizing the moving speed of the decision boundary under Adversarial Training, one of the most effective robust training algorithms, a surprising moving-behavior is revealed: the decision boundary move s away from some vulnerable points but simultaneously moves closer to others, decreasing their margins. To alleviate these conflicting dynamics of the decision boundary, we propose Dynamics-aware Robust Training (DyART), which encourages the decision boundary to engage in movement that prioritizes increasing smaller margins. In contrast to prior works, DyART directly operates on the margins rather than their indirect approximations, allowing for more targeted and effective robustness improvement. Experiments on the CIFAR-10 and Tiny-ImageNet datasets verify that DyART alleviates the conflicting dynamics of the decision boundary and obtains improved robustness under various perturbation sizes compared to the state-of-the-art defenses. Our code is available at https://github.com/Yuancheng-Xu/Dynamics-Aware-Robust-Training.

****

Countering the Attack-Defense Complexity Gap for Robust Classifiers

Samuele Marro,Michele Lombardi

We consider the decision version of defending and attacking Machine Learning classifiers. We provide a rationale for the well-known difficulties in building robust models: in particular we prove that, under broad assumptions, attacking a polynomial-time classifier is NP-complete, while training a polynomial-time model that is robust on even a single input is Σ_2^P-complete. We also provide more general bounds for non-polynomial classifiers. We then show how such a complexity gap can be sidestepped by introducing Counter-Attack (CA), a system that computes on-the-fly robustness certificates for a given input up to an arbitrary distance bound ε. We also prove that, even when attacked with perturbations of magnitude $\varepsilon^\prime > \varepsilon$, CA still provides computational robustness: specifically, while computing a certificate is NP-complete, attacking the system beyond its intended robustness is Σ_2^P-complete. Since the exact form of CA can still be computationally expensive, we introduce a relaxation of this method, which we empirically show to be reliable at identifying non-robust inputs. As part of our work, we introduce UG100, a new dataset obtained by applying a provably optimal attack to six limited-scale networks (three for MNIST and three for CIFAR10), each trained in three different manners.

****

Evaluating Counterfactual Explainers

Diego Velazquez,Pau Rodriguez,Alexandre Lacoste,Issam H. Laradji,Xavier Roca,Jordi Gonzàlez

Explainability methods have been widely used to provide insight into the decisions made by statistical models, thus facilitating their adoption in various domains within the industry. Counterfactual explanation methods aim to improve our understanding of a model by perturbing samples in a way that would alter its response in an unexpected manner. This information is helpful for users and for machine learning practitioners to understand and improve their models. Given the value provided by counterfactual explanations, there is a growing interest in the research community to investigate and propose new methods. However, we identify two issues that could hinder the progress in this field. (1) Existing metrics do not accurately reflect the value of an explainability method for the users. (2) Comparisons between methods are usually performed with datasets like CelebA, where images are annotated with attributes that do not fully describe them and with subjective attributes such as ``Attractive''. In this work, we address these problems by proposing an evaluation method with a principled metric to evaluate and compare different counterfactual explanation methods. The evaluation method is based on a synthetic dataset where images are fully described by their annotated attributes. As a result, we are able to perform a fair comparison of multiple explainability methods in the recent literature, obtaining insights about their performance. We make the code public for the benefit of the research community.

****

SMART: Sentences as Basic Units for Text Evaluation

Reinald Kim Amplayo,Peter J Liu,Yao Zhao,Shashi Narayan

Widely used evaluation metrics for text generation either do not work well with longer texts or fail to evaluate all aspects of text quality. In this paper, we introduce a new metric called SMART to mitigate such limitations. Specifically, we treat sentences as basic units of matching instead of tokens, and use a sentence matching function to soft-match candidate and reference sentences. Candidate sentences are also compared to sentences in the source documents to allow grounding (e.g., factuality) evaluation. Our results show that system-level correlations of our proposed metric with a model-based matching function outperforms all competing metrics on the SummEval summarization meta-evaluation dataset, while the same metric with a string-based matching function is competitive with current model-based metrics. The latter does not use any neural model, which is useful during model development phases where resources can be limited and fast evaluation is required. SMART also outperforms all factuality evaluation metrics on the TRUE benchmark. Finally, we also conducted extensive analyses showing that our proposed metrics work well with longer summaries and are less biased towards specific models.

****

A Reinforcement Learning Approach to Estimating Long-term Treatment Effects

Ziyang Tang,Yiheng Duan,Stephanie Zhang,Lihong Li

Randomized experiments (a.k.a. A/B tests) are a powerful tool for estimating treatment effects, to inform decisions making in business, healthcare and other applications. In many problems, the treatment has a lasting effect that evolves over time. A limitation with randomized experiments is that they do not easily extend to measure long-term effects, since running long experiments is time-consuming and expensive. In this paper, we take a reinforcement learning (RL) approach that estimates the average reward in a Markov process. Motivated by real-world scenarios where the observed state transition is nonstationary, we develop a new algorithm for a class of nonstationary problems, and demonstrate promising results in two synthetic datasets and one online store dataset.

****

Sample-Efficient Reinforcement Learning by Breaking the Replay Ratio Barrier

Pierluca D'Oro,Max Schwarzer,Evgenii Nikishin,Pierre-Luc Bacon,Marc G Bellemare,Aaron Courville

Increasing the replay ratio, the number of updates of an agent's parameters per environment interaction, is an appealing strategy for improving the sample efficiency of deep reinforcement learning algorithms. In this work, we show that fully or partially resetting the parameters of deep reinforcement learning agents causes better replay ratio scaling capabilities to emerge. We push the limits of the sample efficiency of carefully-modified algorithms by training them using an order of magnitude more updates than usual, significantly improving their performance in the Atari 100k and DeepMind Control Suite benchmarks. We then provide an analysis of the design choices required for favorable replay ratio scaling to be possible and discuss inherent limits and tradeoffs.

****

Tier Balancing: Towards Dynamic Fairness over Underlying Causal Factors

Zeyu Tang,Yatong Chen,Yang Liu,Kun Zhang

The pursuit of long-term fairness involves the interplay between decision-making and the underlying data generating process. In this paper, through causal modeling with a directed acyclic graph (DAG) on the decision-distribution interplay, we investigate the possibility of achieving long-term fairness from a dynamic perspective. We propose Tier Balancing, a technically more challenging but more natural notion to achieve in the context of long-term, dynamic fairness analysis. Different from previous fairness notions that are defined purely on observed variables, our notion goes one step further, capturing behind-the-scenes situation changes on the unobserved latent causal factors that directly carry out the influence from the current decision to the future data distribution. Under the specified dynamics, we prove that in general one cannot achieve the long-term fairness goal only through one-step interventions. Furthermore, in the effort of approaching long-term fairness, we consider the mission of "getting closer to" the long-term fairness goal and present possibility and impossibility results accordingly.

****

Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections
Edward De Brouwer,Rahul G Krishnan
Neural ordinary differential equations (Neural ODEs) are an effective framework for learning dynamical systems from irregularly sampled time series data. These models provide a continuous-time latent representation of the underlying dynamical system where new observations at arbitrary time points can be used to update the latent representation of the dynamical system. Existing parameterizations for the dynamics functions of Neural ODEs limit the ability of the model to retain global information about the time series; specifically, a piece-wise integration of the latent process between observations can result in a loss of memory on the dynamic patterns of previously observed data points. We propose PolyODE, a Neural ODE that models the latent continuous-time process as a projection onto a basis of orthogonal polynomials. This formulation enforces long-range memory and preserves a global representation of the underlying dynamical system. Our construction is backed by favourable theoretical guarantees and in a series of experiments, we demonstrate that it outperforms previous works in the reconstruction of past and future data, and in downstream prediction tasks.
****

Neural Design for Genetic Perturbation Experiments
Aldo Pacchiano,Drausin Wulsin,Robert A Barton,Luis Voloch
The problem of how to genetically modify cells in order to maximize a certain cellular phenotype has taken center stage in drug development over the last few years (with, for example, genetically edited CAR-T, CAR-NK, and CAR-NKT cells entering cancer clinical trials). Exhausting the search space for all possible genetic edits (perturbations) or combinations thereof is infeasible due to cost and experimental limitations. This work provides a theoretically sound framework for iteratively exploring the space of perturbations in pooled batches in order to maximize a target phenotype under an experimental budget. Inspired by this application domain, we study the problem of batch query bandit optimization and introduce the Optimistic Arm Elimination (OAE) principle designed to find an almost optimal arm under different functional relationships between the queries (arms) and the outputs (rewards). We analyze the convergence properties of OAE by relating it to the Eluder dimension of the algorithm's function class and validate that OAE outperforms other strategies in finding optimal actions in experiments on simulated problems, public datasets well-studied in bandit contexts, and in genetic perturbation datasets when the regression model is a deep neural network. OAE also outperforms the benchmark algorithms in 3 of 4 datasets in the GeneDisco experimental planning challenge.
****

Conceptual SCAN: Learning With and About Rules
Nathan Scales,Nathanael Schärli,Abubakr Babiker,Yu-Han Liu,Mostafa Dehghani,Olivier Bousquet
The ability to learn from a mix of rules and examples and to reflect on the learned abstractions is an important aspect of human intelligence. At the same time, there is a lack of benchmarks that systematically test for this ability, which makes it hard to evaluate the degree to which it is present in state-of-the-art ML architectures. We introduce a method to systematically construct such benchmarks by using an example structure that allows us to explicitly provide and ask about rules that are relevant for the given task. We present a simple dataset that is constructed according to this method, and we use it to analyze the performance of a variety of T5-based machine learning models. We identify four challenge areas in this setup: maintaining consistency between learned rules and their application, scaling to larger rule sets, compositional generalization, and dealing with limited training data.
****

Have Missing Data? Make It Miss More! Imputing Tabular Data with Masked Autoencoding
Tianyu Du,Luca Melis,Ting Wang
We present ReMasker, a novel method for imputing missing values in tabular data by extending the masked autoencoding framework. In contrast to prior work, ReMas

ker is both {\em simple} -- besides the missing values (i.e., naturally masked), we randomly ``re-mask'' another set of values, optimize the autoencoder by reconstructing this re-masked set, and apply the trained model to predict the missing values; and {\em effective} -- with extensive evaluation on benchmark datasets, we show that ReMasker consistently outperforms state-of-the-art methods in terms of both imputation fidelity and utility under various missingness settings, while its performance advantage often increases with the ratio of missing data. We further explore theoretical justification for its effectiveness, showing that ReMasker tends to learn missingness-invariant representations of tabular data. Our findings indicate that masked modeling represents a promising direction for further research on tabular data imputation.

****

Invertible normalizing flow neural networks by JKO scheme
Chen Xu,Xiuyuan Cheng,Yao Xie

Normalizing flow is a class of deep generative models for efficient sampling and density estimation. In practice, the flow often appears as a chain of invertible neural network blocks. To facilitate training, past works have regularized flow trajectories and designed special network architectures. The current paper develops a neural ODE flow network inspired by the Jordan-Kinderleherer-Otto (JKO) scheme, which allows an efficient \textit{block-wise} training procedure: as the JKO scheme unfolds the dynamic of gradient flow, the proposed model naturally stacks residual network blocks one-by-one and reduces the memory load as well as the difficulty of training deep networks. We also develop an adaptive time-reparametrization of the flow network with a progressive refinement of the trajectory in probability space, which improves the optimization efficiency and model accuracy in practice.

On high-dimensional generative tasks for tabular data, JKO-Flow can process larger data batches and perform competitively as or better than continuous and discrete flow models, using 10X less number of iterations (e.g., batches) and significantly less time per iteration.

****

Unsupervised learning of features and object boundaries from local prediction
Heiko H. Schütt,Wei Ji Ma

A visual system has to learn both which features to extract from images and how to group locations into (proto-)objects. Those two aspects are usually dealt with separately, although predictability is discussed as a cue for both. To incorporate features and boundaries into the same model, we model a layer of feature maps with a pairwise Markov random field model in which each factor is paired with an additional binary variable, which switches the factor on or off. Using one of two contrastive learning objectives, we can learn both the features and the parameters of the Markov random field factors from images without further supervision signals. The features learned by shallow neural networks based on this loss are local averages, opponent colors, and Gabor-like stripe patterns. Furthermore, we can infer connectivity between locations by inferring the switch variables. Contours inferred from this connectivity perform quite well on the Berkeley segmentation database (BSDS500) without any training on contours. Thus, computing predictions across space aids both segmentation and feature learning, and models trained to optimize these predictions show similarities to the human visual system. We speculate that retinotopic visual cortex might implement such predictions over space through lateral connections.

****

Multi-Segmental Informational Coding for Self-Supervised Representation Learning
Chuang Niu,Ge Wang

Self-supervised representation learning aims to map high-dimensional data into a compact embedding space, where samples with similar semantics are close to each other. Currently, most representation learning methods maximize the cosine similarity or minimize the distance between different views from the same sample in an ℓ^2 normalized embedding space, and reduce the feature redundancy via a linear correlation constraint. In this study, we propose MUlti-Segmental Informational Coding (MUSIC) as a new embedding scheme for self-supervised representati on learning. MUSIC divides an embedding vector into multiple segments to represent different types of attributes, and each segment automatically learns a set of discrete and complementary attributes. MUSIC enables the estimation of the probability distribution over discrete attributes and thus the learning process can be directly guided by information measurements, reducing the feature redundancy beyond the linear correlation. Our theoretical analysis guarantees that MUSIC learns transform-invariant, non-trivial, diverse, and discriminative features. MUSIC does not require a special asymmetry design, a very high dimension of embedding features, or a deep projection head, making the training framework flexible and efficient. Extensive experiments demonstrate the superiority of MUSIC.

****

Rule-based policy regularization for reinforcement learning-based building control

Hsin Yu Liu,Bharathan Balaji,Rajesh Gupta,Dezhi Hong

Rule-based control (RBC) is widely adopted in buildings due to its stability and robustness. It resembles a behavior cloning methodology refined by human expertise. However, it is unlikely for RBC to exceed a reinforcement learning (RL) agent’s performance since it is challenging to ingest a large number of parameters during decision-making. In this paper, we explore how to incorporate rule-based control into reinforcement learning to learn a more robust policy in both online and offline settings with a unified approach. We start with state-of-the-art online and offline RL methods, TD3 and TD3+BC, then improve on them using a dynamically weighted actor loss function to selectively choose which policy should RL models learn from at each time step of training. With experiments across multiple tasks and various weather conditions in both deterministic and stochastic scenarios, we empirically demonstrate that our dynamically weighted rule-based incorporated control regularization (RUBICON) method outperforms representative baseline methods in offline settings by 40.7% in a reward settings consisting of the combination of thermal comfort and energy consumption and by 49.7% in online settings in building-RL environments.

****

Neural Graphical Models

Harsh Shrivastava,Urszula Chajewska

Graphs are ubiquitous and are often used to understand the dynamics of a system. Probabilistic Graphical Models comprising Bayesian and Markov networks, and Conditional Independence graphs are some of the popular graph representation techniques. They can model relationships between features (nodes) together with the underlying distribution. Although theoretically these models can represent very complex dependency functions, in practice often simplifying assumptions are made due to computational limitations associated with graph operations. This work introduces Neural Graphical Models (NGMs) which attempt to represent complex feature dependencies with reasonable computational costs. Specifically, given a graph of feature relationships and corresponding samples, we capture the dependency structure between the features along with their complex function representations by using neural networks as a multi-task learning framework. We provide efficient learning, inference and sampling algorithms for NGMs. Moreover, NGMs can fit generic graph structures including directed, undirected and mixed-edge graphs as well as support mixed input data types. We present empirical studies that show NGMs' capability to represent Gaussian graphical models, inference analysis of a lung cancer data and extract insights from a real world infant mortality data provided by CDC.

****

AUGMENTING ZERO-SHOT DENSE RETRIEVERS WITH PLUG-IN MIXTURE-OF-MEMORIES

Suyu Ge,Chenyan Xiong,Corby Louis Rosset,Arnold Overwijk,Jiawei Han,Paul N. Bennett

In this paper we improve the zero-shot generalization ability of language models via Mixture-Of-Memory Augmentation (MoMA), a mechanism that retrieves augmentation documents from multiple information corpora (“external memories”), with the option to “plug in” new memory at inference time. We develop a joint learning mechanism that trains the augmentation component with latent labels derived from t he end retrieval task, paired with hard negatives from the memory mixture. We instantiate the model in a zero-shot dense retrieval setting by augmenting a strong T5-based retriever with MoMA. Our model, MoMA-DR, obtains strong zero-shot retrieval accuracy on the eighteen tasks included in the standard BEIR benchmark. It outperforms other dense retrieval models of similar scales and achieves comparable accuracy with systems that seek generalization from increased scales in encoder models or vector indices. Our analysis illustrates the necessity of augmenting with mixture-of-memory for robust generalization, the benefits of joint learning, and how MoMA-DR utilizes the plug-in memory at inference time without changing its parameters. We plan to open source our code.

****

Efficient Discrete Multi Marginal Optimal Transport Regularization

Ronak Mehta,Jeffery Kline,Vishnu Suresh Lokhande,Glenn Fung,Vikas Singh

Optimal transport has emerged as a powerful tool for a variety of problems in machine learning, and it is frequently used to enforce distributional constraints. In this context, existing methods often use either a Wasserstein metric, or else they apply concurrent barycenter approaches when more than two distributions are considered. In this paper, we leverage multi-marginal optimal transport (MMOT), where we take advantage of a procedure that computes a generalized earth mover's distance as a sub-routine. We show that not only is our algorithm computationally more efficient compared to other barycentric-based distance methods, but it has the additional advantage that gradients used for backpropagation can be efficiently computed during the forward pass computation itself, which leads to substantially faster model training. We provide technical details about this new regularization term and its properties, and we present experimental demonstrations of faster runtimes when compared to standard Wasserstein-style methods. Finally, on a range of experiments designed to assess effectiveness at enforcing fairness, we demonstrate our method compares well with alternatives.

****

AutoTransfer: AutoML with Knowledge Transfer - An Application to Graph Neural Networks

Kaidi Cao,Jiaxuan You,Jiaju Liu,Jure Leskovec

AutoML has demonstrated remarkable success in finding an effective neural architecture for a given machine learning task defined by a specific dataset and an evaluation metric. However, most present AutoML techniques consider each task independently from scratch, which requires exploring many architectures, leading to high computational cost. Here we propose AutoTransfer, an AutoML solution that improves search efficiency by transferring the prior architectural design knowledge to the novel task of interest. Our key innovation includes a task-model bank that captures the model performance over a diverse set of GNN architectures and tasks, and a computationally efficient task embedding that can accurately measure the similarity among different tasks. Based on the task-model bank and the task embeddings, we estimate the design priors of desirable models of the novel task, by aggregating a similarity-weighted sum of the top-K design distributions on tasks that are similar to the task of interest. The computed design priors can be used with any AutoML search algorithm. We evaluate AutoTransfer on six datasets in the graph machine learning domain. Experiments demonstrate that (i) our proposed task embedding can be computed efficiently, and that tasks with similar embeddings have similar best-performing architectures; (ii) AutoTransfer significantly improves search efficiency with the transferred design priors, reducing the number of explored architectures by an order of magnitude. Finally, we release GNN-Bank-101, a large-scale dataset of detailed GNN training information of 120,000 task-model combinations to facilitate and inspire future research.

****

Meta-learning from demonstrations improves compositional generalization

Sam Spilsbury,Alexander Ilin

We study the problem of compositional generalization of language-instructed agents in gSCAN. gSCAN is a popular benchmark which requires an agent to generalize to instructions containing novel combinations of words, which are not seen in the training data. We propose to improve the agent’s generalization capabilities w ith an architecture inspired by the Meta-Sequence-to-Sequence learning approach (Lake, 2019). The agent receives as a context a few examples of pairs of instructions and action trajectories in a given instance of the environment (a support set) and it is tasked to predict an action sequence for a query instruction for the same environment instance. The context is generated by an oracle and the instructions come from the same distribution as seen in the training data. In each training episode, we also shuffle the indices of the actions and the words of the instructions to make the agent figure out the relations between the actions and the words from the context. Our predictive model has the standard transformer architecture. We show that the proposed architecture can significantly improve the generalization capabilities of the agent on one of the most difficult gSCAN splits: the ``adverb-to-verb” split H.

****

Deep Dependency Networks for Action Classification in Video

Shivvrat Arya,Yu Xiang,Vibhav Giridhar Gogate

We propose a simple approach which combines the strengths of probabilistic graphical models and deep learning architectures for solving the multi-label action classification task in videos. At a high level, given a video clip, the goal in this task is to infer the set of activities, defined as verb-noun pairs, that are performed in the clip. First, we show that the performance of previous approaches that combine Markov Random Fields with neural networks can be modestly improved by leveraging more powerful methods such as iterative join graph propagation, ℓ-1 regularization based structure learning and integer linear programming. Then we propose a new modeling framework called deep dependency network which augments a dependency network, a model that is easy to train and learns more accurate dependencies but is limited to Gibbs sampling for inference, to the output layer of a neural network. We show that despite its simplicity, joint learning this new architecture yields significant improvements in performance over the baseline neural network. In particular, our experimental evaluation on three video datasets: Charades, Textually Annotated Cooking Scenes (TaCOS), and Wetlab shows that deep dependency networks are almost always superior to pure neural architectures that do not use dependency networks.

****

Temporal Dependencies in Feature Importance for Time Series Prediction

Kin Kwan Leung,Clayton Rooke,Jonathan Smith,Saba Zuberi,Maksims Volkovs

Time series data introduces two key challenges for explainability methods: firstly, observations of the same feature over subsequent time steps are not independent, and secondly, the same feature can have varying importance to model predictions over time. In this paper, we propose Windowed Feature Importance in Time (WinIT), a feature removal based explainability approach to address these issues. Unlike existing feature removal explanation methods, WinIT explicitly accounts for the temporal dependence between different observations of the same feature in the construction of its importance score. Furthermore, WinIT captures the varying importance of a feature over time, by summarizing its importance over a window of past time steps. We conduct an extensive empirical study on synthetic and real-world data, compare against a wide range of leading explainability methods, and explore the impact of various evaluation strategies. Our results show that WinIT achieves significant gains over existing methods, with more consistent performance across different evaluation metrics.

****

Peaks2Image: Reconstructing fMRI Statistical Maps from Peaks

Raphaël Meudec,Jérôme Dockès,Demian Wassermann,Bertrand Thirion

Neuroscience is striving to overcome the lack of power due to the small sample size of standard studies. An important step forward has been the creation of large-scale public image repositories, such as NeuroVault. Such repositories enable comparing images across studies and automatically associating them with cognitive terms. Yet, this type of meta-analysis faces a major roadblock: the scarcity and inconsistency of image annotations and metadata. Another resource containing rich annotations is the neuroscientific literature. However it only yields a handful of brain-space coordinates per publication, those of the main activity peak s reported in each study. In this work, we propose Peaks2Image, a neuralnetwork approach to reconstructing continuous spatial representations of brain activity from peak activation tables. Using reconstructions of studies published in the neuroscientific literature, we train a decoder using tf-idf features as labels, leading to a much broader set of decoded terms than current image-based studies. We validate the decoder on 43,000 NeuroVault images, successfully decoding 58 out of 81 concepts in a zero-shot setting.

****

Bridging the Gap between Semi-supervised and Supervised Continual Learning via Data Programming

Pengyuan Lu,Seungwon Lee,Amanda Annette Watson,David Kent,Insup Lee,ERIC EATON,James Weimer

Semi-supervised continual learning (SSCL) has shown its utility in learning cumulative knowledge with partially labeled data per task. However, the state-of-the-art has yet to explicitly address how to reduce the performance gap between using partially labeled data and fully labeled. In response, we propose a general-purpose SSCL framework, namely DP-SSCL, that uses data programming (DP) to pseudo-label the unlabeled data per task, and then cascades both ground-truth-labeled and pseudo-labeled data to update a downstream supervised continual learning model. The framework includes a feedback loop that brings mutual benefits: On one hand, DP-SSCL inherits guaranteed pseudo-labeling quality from DP techniques to improve continual learning, approaching the performance of using fully supervised data. On the other hand, knowledge transfer from previous tasks facilitates training of the DP pseudo-labeler, taking advantage of cumulative information via self-teaching. Experiments show that (1) DP-SSCL bridges the performance gap, approaching the final accuracy and catastrophic forgetting as using fully labeled data, (2) DP-SSCL outperforms existing SSCL approaches at low cost, by up to 25% higher final accuracy and lower catastrophic forgetting on standard benchmarks, while reducing memory overhead from 100 MB level to 1 MB level at the same time complexity, and (3) DP-SSCL is flexible, maintaining steady performance supporting plug-and-play extensions for a variety of supervised continual learning models.

****

Characterizing the spectrum of the NTK via a power series expansion

Michael Murray,Hui Jin,Benjamin Bowman,Guido Montufar

Under mild conditions on the network initialization we derive a power series expansion for the Neural Tangent Kernel (NTK) of arbitrarily deep feedforward networks in the infinite width limit. We provide expressions for the coefficients of this power series which depend on both the Hermite coefficients of the activation function as well as the depth of the network. We observe faster decay of the Hermite coefficients leads to faster decay in the NTK coefficients and explore the role of depth. Using this series, first we relate the effective rank of the NTK to the effective rank of the input-data Gram. Second, for data drawn uniformly on the sphere we study the eigenvalues of the NTK, analyzing the impact of the choice of activation function. Finally, for generic data and activation functions with sufficiently fast Hermite coefficient decay, we derive an asymptotic upper bound on the spectrum of the NTK.

****

Unmasking the Lottery Ticket Hypothesis: What's Encoded in a Winning Ticket's Mask?

Mansheej Paul,Feng Chen,Brett W. Larsen,Jonathan Frankle,Surya Ganguli,Gintare Karolina Dziugaite

Modern deep learning involves training costly, highly overparameterized networks, thus motivating the search for sparser networks that require less compute and memory but can still be trained to the same accuracy as the full network (i.e. matching). Iterative magnitude pruning (IMP) is a state of the art algorithm that can find such highly sparse matching subnetworks, known as winning tickets. IMP operates by iterative cycles of training, masking a fraction of smallest magnitude weights, rewinding unmasked weights back to an early training point, and repeating. Despite its simplicity, the underlying principles for when and how IMP f inds winning tickets remain elusive. In particular, what useful information does an IMP mask found at the end of training convey to a rewound network near the beginning of training? How does SGD allow the network to extract this information? And why is iterative pruning needed, i.e. why can't we prune to very high sparsities in one shot? We develop answers to these questions in terms of the geometry of the error landscape. First, we find that—at higher sparsities—pairs of pruned networks at successive pruning iterations are connected by a linear path with zero error barrier if and only if they are matching. This indicates that masks found at the end of training convey to the rewind point the identity of an axial subspace that intersects a desired linearly connected mode of a matching sublevel set. Second, we show SGD can exploit this information due to a strong form of robustness: it can return to this mode despite strong perturbations early in training. Third, we show how the flatness of the error landscape at the end of training determines a limit on the fraction of weights that can be pruned at each iteration of IMP. This analysis yields a new quantitative link between IMP performance and the Hessian eigenspectrum. Finally, we show that the role of retraining in IMP is to find a network with new small weights to prune. Overall, these results make progress toward demystifying the existence of winning tickets by revealing the fundamental role of error landscape geometry in the algorithms used to find them.

****

A critical look at the evaluation of GNNs under heterophily: Are we really making progress?

Oleg Platonov,Denis Kuznedelev,Michael Diskin,Artem Babenko,Liudmila Prokhorenkova

Node classification is a classical graph representation learning task on which Graph Neural Networks (GNNs) have recently achieved strong results. However, it is often believed that standard GNNs only work well for homophilous graphs, i.e., graphs where edges tend to connect nodes of the same class. Graphs without this property are called heterophilous, and it is typically assumed that specialized methods are required to achieve strong performance on such graphs. In this work, we challenge this assumption. First, we show that the standard datasets used for evaluating heterophily-specific models have serious drawbacks, making results obtained by using them unreliable. The most significant of these drawbacks is the presence of a large number of duplicate nodes in the datasets Squirrel and Chameleon, which leads to train-test data leakage. We show that removing duplicate nodes strongly affects GNN performance on these datasets. Then, we propose a set of heterophilous graphs of varying properties that we believe can serve as a better benchmark for evaluating the performance of GNNs under heterophily. We show that standard GNNs achieve strong results on these heterophilous graphs, almost always outperforming specialized models. Our datasets and the code for reproducing our experiments are available at https://github.com/yandex-research/heterophilous-graphs

****

Dr.Spider: A Diagnostic Evaluation Benchmark towards Text-to-SQL Robustness

Shuaichen Chang,Jun Wang,Mingwen Dong,Lin Pan,Henghui Zhu,Alexander Hanbo Li,Wuwei Lan,Sheng Zhang,Jiarong Jiang,Joseph Lilien,Steve Ash,William Yang Wang,Zhiguo Wang,Vittorio Castelli,Patrick Ng,Bing Xiang

Neural text-to-SQL models have achieved remarkable performance in translating natural language questions into SQL queries. However, recent studies reveal that text-to-SQL models are vulnerable to task-specific perturbations. Previous curated robustness test sets usually focus on individual phenomena. In this paper, we propose a comprehensive robustness benchmark based on Spider, a cross-domain text-to-SQL benchmark, to diagnose the model robustness. We design 17 perturbations on databases, natural language questions, and SQL queries to measure the robustness from different angles. In order to collect more diversified natural question perturbations, we utilize large pretrained language models (PLMs) to simulate human behaviors in creating natural questions. We conduct a diagnostic study of the state-of-the-art models on the robustness set. Experimental results reveal that even the most robust model suffers from a 14.0% performance drop overall and a 50.7% performance drop on the most challenging perturbation. We also present a breakdown analysis regarding text-to-SQL model designs and provide insights for improving model robustness.

****

A Non-monotonic Self-terminating Language Model

Eugene Choi,Kyunghyun Cho,Cheolhyoung Lee

Recent large-scale neural autoregressive sequence models have shown impressive performances on a variety of natural language generation tasks. However, their generated sequences often exhibit degenerate properties such as non-termination, undesirable repetition, and premature termination, when generated with decoding algorithms such as greedy search, beam search, top-k sampling, and nucleus sampling. In this paper, we focus on the problem of non-terminating sequences resulting from an incomplete decoding algorithm. We first define an incomplete probable decoding algorithm which includes greedy search, top-k sampling, and nucleus sampling, beyond the incomplete decoding algorithm originally put forward by Welleck et al. (2020). We then propose a non-monotonic self-terminating language model, which significantly relaxes the constraint of monotonically increasing termination probability in the originally proposed self-terminating language model by Welleck et al. (2020), to address the issue of non-terminating sequences when using incomplete probable decoding algorithms. We prove that our proposed model prevents non-terminating sequences when using not only incomplete probable decoding algorithms but also beam search. We empirically validate our model on sequence completion tasks with various architectures.

****

uGLAD: A deep learning model to recover conditional independence graphs

Harsh Shrivastava,Urszula Chajewska,Robin Abraham,Xinshi Chen

Probabilistic Graphical Models are generative models of complex systems. They rely on conditional independence assumptions between variables to learn sparse representations which can be visualized in a form of a graph. Such models are used for domain exploration and structure discovery in poorly understood domains. This work introduces a novel technique to perform sparse graph recovery by optimizing deep unrolled networks. Assuming that the input data $X\in\mathbb{R}^{M\times D}$ comes from an underlying multivariate Gaussian distribution, we apply a deep model on X that outputs the precision matrix Θ. Then, the partial correlation matrix \mathrm{P} is calculated which can also be interpreted as providing a list of conditional independence assertions holding in the input distribution. Our model, \texttt{uGLAD}, builds upon and extends the state-of-the-art model \texttt{GLAD} to the unsupervised setting. The key benefits of our model are (1) \texttt{uGLAD} automatically optimizes sparsity-related regularization parameters leading to better performance than existing algorithms. (2) We introduce multi-task learning based `consensus' strategy for robust handling of missing data in an unsupervised setting. We evaluate performance on synthetic Gaussian, non-Gaussian data generated from Gene Regulatory Networks, and present case studies in anaerobic digestion and infant mortality.

****

Quantifying Memorization Across Neural Language Models

Nicholas Carlini,Daphne Ippolito,Matthew Jagielski,Katherine Lee,Florian Tramer,Chiyuan Zhang

Large language models (LMs) have been shown to memorize parts of their training data, and when prompted appropriately, they will emit the memorized training data verbatim. This is undesirable because memorization violates privacy (exposing user data), degrades utility (repeated easy-to-memorize text is often low quality), and hurts fairness (some texts are memorized over others).

We describe three log-linear relationships that quantify the degree to which LMs emit memorized training data. Memorization significantly grows as we increase (1) the capacity of a model, (2) the number of times an example has been duplicated, and (3) the number of tokens of context used to prompt the model. Surprisingly, we find the situation becomes complicated when generalizing these results across model families. On the whole, we find that memorization in LMs is more prevalent than previously believed and will likely get worse as models continues to scale, at least without active mitigations.

****

Powderworld: A Platform for Understanding Generalization via Rich Task Distributions

Kevin Frans,Phillip Isola

One of the grand challenges of reinforcement learning is the ability to generalize to new tasks. However, general agents require a set of rich, diverse tasks to train on. Designing a `foundation environment' for such tasks is tricky -- the ideal environment would support a range of emergent phenomena, an expressive task space, and fast runtime. To take a step towards addressing this research bottleneck, this work presents Powderworld, a lightweight yet expressive simulation environment running directly on the GPU. Within Powderworld, two motivating task distributions are presented, one for world-modelling and one for reinforcement learning. Each contains hand-designed test tasks to examine generalization. Experiments indicate that increasing the environment's complexity improves generalization for world models, yet causes reinforcement learning agents to struggle. Powderworld aims to support the study of generalization by providing a source of diverse tasks arising from the same core rules.

****

Federated Self-supervised Learning for Heterogeneous Clients

Disha Makhija,Nhat Ho,Joydeep Ghosh

Federated Learning has become an important learning paradigm due to its privacy and computational benefits. As the field advances, two key challenges that still remain to be addressed are: (1) system heterogeneity - variability in the compute and/or data resources present on each client, and (2) lack of labeled data in certain federated settings. Several recent developments have tried to overcome these challenges independently. In this work, we propose a unified and systematic framework, \emph{Heterogeneous Self-supervised Federated Learning} (Hetero-SSFL) for enabling self-supervised learning with federation on heterogeneous clients. The proposed framework allows collaborative representation learning across all the clients without imposing architectural constraints or requiring presence of labeled data. The key idea in Hetero-SSFL is to let each client train its unique self-supervised model and enable the joint learning across clients by aligning the lower dimensional representations on a common dataset. The entire training procedure could be viewed as self and peer-supervised as both the local training and the alignment procedures do not require presence of any labeled data. As in conventional self-supervised learning, the obtained client models are task independent and can be used for varied end-tasks.

We provide a convergence guarantee of the proposed framework for non-convex objectives in heterogeneous settings and also empirically demonstrate that our proposed approach outperforms the state of the art methods by a significant margin.

****

ContraSim -- A Similarity Measure Based on Contrastive Learning

Adir Rahamim,Yonatan Belinkov

Recent work has compared neural network representations via similarity-based analyses, shedding light on how different aspects (architecture, training data, etc.) affect models' internal representations. The quality of a similarity measure is typically evaluated by its success in assigning a high score to representations that are expected to be matched. However, existing similarity measures perform mediocrely on standard benchmarks. In this work, we develop a new similarity measure, dubbed ContraSim, based on contrastive learning. In contrast to common closed-form similarity measures, ContraSim learns a parameterized measure by using both similar and dissimilar examples. We perform an extensive experimental evaluation of our method, with both language and vision models, on the standard layer prediction benchmark and two new benchmarks that we develop: the multilingual benchmark and the image--caption benchmark. In all cases, ContraSim achieves much higher accuracy than previous similarity measures, even when presented with challenging examples.

****

Learning to Segment from Noisy Annotations: A Spatial Correction Approach

Jiachen Yao,Yikai Zhang,Songzhu Zheng,Mayank Goswami,Prateek Prasanna,Chao Chen

Noisy labels can significantly affect the performance of deep neural networks (DNNs). In medical image segmentation tasks, annotations are error-prone due to the high demand in annotation time and in the annotators' expertise. Existing methods mostly tackle label noise in classification tasks. Their independent-noise assumptions do not fit label noise in segmentation task. In this paper, we propose a novel noise model for segmentation problems that encodes spatial correlation and bias, which are prominent in segmentation annotations. Further, to mitigate such label noise, we propose a label correction method to recover true label progressively. We provide theoretical guarantees of the correctness of the proposed method. Experiments show that our approach outperforms current state-of-the-art methods on both synthetic and real-world noisy annotations.

****

Measuring Forgetting of Memorized Training Examples

Matthew Jagielski,Om Thakkar,Florian Tramer,Daphne Ippolito,Katherine Lee,Nicholas Carlini,Eric Wallace,Shuang Song,Abhradeep Guha Thakurta,Nicolas Papernot,Chiyuan Zhang

Machine learning models exhibit two seemingly contradictory phenomena: training data memorization and various forms of forgetting. In memorization, models overfit specific training examples and become susceptible to privacy attacks. In forgetting, examples which appeared early in training are forgotten by the end. In this work, we connect these phenomena.

We propose a technique to measure to what extent models ``forget'' the specifics of training examples, becoming less susceptible to privacy attacks on examples they have not seen recently.

We show that, while non-convexity can prevent forgetting from happening in the worst-case, standard image,speech, and language models empirically do forget examples over time.

We identify nondeterminism as a potential explanation, showing that deterministically trained models do not forget.

Our results suggest that examples seen early when training with extremely large datasets---for instance those examples used to pre-train a model---may observe privacy benefits at the expense of examples seen later.

****

Graduated Non-Convexity for Robust Self-Trained Language Understanding

Hongyin Luo,Yoon Kim,James R. Glass

Self-training has been proved an efficient strategy for unsupervised fine-tuning of language models using unlabeled data and model-generated pseudo-labels. However, the performance of self-trained models is unstable under different settings of training and evaluation data, influenced by both data distribution and pseudo-label accuracy. In this work, we propose an outlier robust self-training method based on graduated non-convexity (GNC) to mitigate the problem. We construct self-training as a non-convex optimization problem with outlier training examples. The models are self-trained with robust cost functions based according to Black-Rangarajan Duality. The algorithm learns slack variables as the loss weights for all training samples. The slack variables are used to calibrate the loss items during training to update the model parameters. The calibrated loss items lead to more robust self-trained models against different training and evaluation data and tasks. We conducted experiments on few-shot natural language understanding tasks with labeled and unlabeled data examples. Experiment results show that the proposed loss calibration method improves the performance and stability of self-training under different training tasks and data examples, and also benefits the robustness against adversarial evaluation corpora.

****

MaskViT: Masked Visual Pre-Training for Video Prediction

Agrim Gupta,Stephen Tian,Yunzhi Zhang,Jiajun Wu,Roberto Martín-Martín,Li Fei-Fei

The ability to predict future visual observations conditioned on past observations and motor commands can enable embodied agents to plan solutions to a variety of tasks in complex environments. This work shows that we can create good video prediction models by pre-training transformers via masked visual modeling. Our a pproach, named MaskViT, is based on two simple design decisions. First, for memory and training efficiency, we use two types of window attention: spatial and spatiotemporal. Second, during training, we mask a variable percentage of tokens instead of a fixed mask ratio. For inference, MaskViT generates all tokens via iterative refinement where we incrementally decrease the masking ratio following a mask scheduling function. On several datasets we demonstrate that MaskViT outperforms prior works in video prediction, is parameter efficient, and can generate high resolution videos ($256 \times $256). Further, we demonstrate the benefits of inference speedup (up to $512 \times$) due to iterative decoding by using MaskViT for planning on a real robot. Our work suggests that we can endow embodied agents with powerful predictive models by leveraging the general framework of masked visual modeling with minimal domain knowledge.

****

Text Summarization with Oracle Expectation
Yumo Xu,Mirella Lapata

Extractive summarization produces summaries by identifying and concatenating the most important sentences in a document. Since most summarization datasets do not come with gold labels indicating whether document sentences are summary-worthy, different labeling algorithms have been proposed to extrapolate oracle extracts for model training. In this work, we identify two flaws with the widely used greedy labeling approach: it delivers suboptimal and deterministic oracles. To alleviate both issues, we propose a simple yet effective labeling algorithm that creates soft, expectation-based sentence labels. We define a new learning objective for extractive summarization which incorporates learning signals from multiple oracle summaries and prove it is equivalent to estimating the oracle expectation for each document sentence. Without any architectural modifications, the proposed labeling scheme achieves superior performance on a variety of summarization benchmarks across domains and languages, in both supervised and zero-shot settings.

****

MERMADE: K-shot Robust Adaptive Mechanism Design via Model-Based Meta-Learning
Arundhati Banerjee,Soham Rajesh Phade,Stefano Ermon,Stephan Zheng

Mechanism design (MD) studies how rules and rewards shape the behavior of intelligent agents, e.g., in auctions or the economy. Simulations with AI agents are powerful tools for MD, but real-world agents may behave and learn differently than simulated agents under a given mechanism. Also, the mechanism designer may not fully observe an agent's learning strategy or rewards, and executing a mechanism may be costly, e.g., enforcing a tax might require extra labor. Hence, it is key to design robust adaptive mechanisms that generalize well to agents with unseen (learning) behavior, are few-shot adaptable, and are cost-efficient. Here, we introduce MERMADE, a model-based meta-learning framework to learn mechanisms that can quickly adapt when facing out-of-distribution agents with different learning strategies and reward functions. First, we show that meta-learning allows adapting to the theoretically known and appropriate Stackelberg equilibrium in a simple matrix game at meta-test time, with few interactions with the agent. Second, with bandit agents, we show empirically that our approach yields strong meta-test time performance against agents with various unseen explore-exploit behaviors. Finally, we outperform baselines that separately use either meta-learning or agent behavior modeling to learn a cost-effective mechanism that is K-shot adaptable with only partial information about the agents.

****

Continuous-time identification of dynamic state-space models by deep subspace encoding
Gerben I. Beintema,Maarten Schoukens,Roland Tóth

Continuous-time (CT) modeling has proven to provide improved sample efficiency and interpretability in learning the dynamical behavior of physical systems compared to discrete-time (DT) models. However, even with numerous recent developments, the CT nonlinear state-space (NL-SS) model identification problem remains to be solved in full, considering common experimental aspects such as the presence of external inputs, measurement noise, latent states, and general robustness. Th is paper presents a novel estimation method that addresses all these aspects and that can obtain state-of-the-art results on multiple benchmarks with compact fully connected neural networks capturing the CT dynamics. The proposed estimation method called the subspace encoder approach (SUBNET) ascertains these results by efficiently approximating the complete simulation loss by evaluating short simulations on subsections of the data, by using an encoder function to estimate the initial state for each subsection and a novel state-derivative normalization to ensure stability and good numerical conditioning of the training process. We prove that the use of subsections increases cost function smoothness together with the necessary requirements for the existence of the encoder function and we show that the proposed state-derivative normalization is essential for reliable estimation of CT NL-SS models.

****

Waveformer: Linear-Time Attention with Forward and Backward Wavelet Transform

Yufan Zhuang,Zihan Wang,Fangbo Tao,Jingbo Shang

We propose Waveformer that learns attention mechanism in the wavelet coefficient space, requires only linear time complexity, and enjoys universal approximating power. Specifically, we first apply forward wavelet transform to project the input sequences to multi-resolution orthogonal wavelet bases, then conduct nonlinear transformations (in this case, a random feature kernel) in the wavelet coefficient space, and finally reconstruct the representation in input space via backward wavelet transform. We note that other non-linear transformations may be used, hence we name the learning paradigm Wavelet transformatIon for Sequence lEarning (WISE). We emphasize the importance of backward reconstruction in the WISE paradigm — without it, one would be mixing information from both the input space and coefficient space through skip connections, which shall not be considered as mathematically sound. Compared with Fourier transform in recent works, wavelet transform is more efficient in time complexity and better captures local and positional information; we further support this through our ablation studies. Extensive experiments on seven long-range understanding datasets from the Long Range Arena benchmark and code understanding tasks demonstrate that (1) Waveformer achieves competitive and even better accuracy than a number of state-of-the-art Transformer variants and (2) WISE can boost accuracies of various attention approximation methods without increasing the time complexity. These together showcase the superiority of learning attention in a wavelet coefficient space over the input space.

****

SaMoE: Parameter Efficient MoE Language Models via Self-Adaptive Expert Combination

Minjia Zhang,Conglong Li,Xiaoxia Wu,Zhewei Yao,Yuxiong He

Recently, Mixture-of-Experts (MoE) has demonstrated success in scaling models to have large amounts of parameters without significant increases in computational cost. However, MoEs have been also reported to be parameter inefficient such that larger models do not always lead to better performance.

In this work, we study how to build parameter-efficient MoE models. Our analysis identifies that MoE layers exhibit poor gradient flow as the number of experts increases, leading to insufficient training of experts. To overcome this issue, we propose a new MoE architecture design (SaMoE), which improves the parameter efficiency of MoE models by learning a soft combination of a global set of expert layers for each MoE layer. Such a scheme enables substantial parameter savings on MoE while achieving comparable or better accuracy than the standard MoE training baseline. Extensive experiments on billion-scale GPT-3 style autoregressive MoE language models demonstrate that SaMoE significantly improves the parameter efficiency of MoE models by reducing up to 5.2X total parameters while obtaining superior pre-training and zero-shot generalization results as compared to baseline.

****

How to Train your HIPPO: State Space Models with Generalized Orthogonal Basis Projections

Albert Gu,Isys Johnson,Aman Timalsina,Atri Rudra,Christopher Re

Linear time-invariant state space models (SSM) are a classical model from engineering and statistics, that have recently been shown to be very promising in machine learning through the Structured State Space sequence model (S4). A core component of S4 involves initializing the SSM state matrix to a particular matrix called a HiPPO matrix, which was empirically important for S4's ability to handle long sequences. However, the specific matrix that S4 uses was actually derived in previous work for a particular *time-varying* dynamical system, and the use of this matrix as a *time-invariant* SSM had no known mathematical interpretation. Consequently, the theoretical mechanism by which S4 models long-range dependencies actually remains unexplained. We derive a more general and intuitive formulation of the HiPPO framework, which provides a simple mathematical interpretation of S4 as a decomposition onto exponentially-warped Legendre polynomials, explaining its ability to capture long dependencies. Our generalization introduces a theoretically rich class of SSMs that also lets us derive more intuitive S4 variants for other bases such as the Fourier basis, and explains other aspects of training S4, such as how to initialize the important timescale parameter. These insights improve S4's performance to 86% on the Long Range Arena benchmark, with 96% on the most difficult Path-X task.

****

Interpretable Debiasing of Vectorized Language Representations with Iterative Orthogonalization

Prince Osei Aboagye,Yan Zheng,Jack Shunn,Chin-Chia Michael Yeh,Junpeng Wang,Zhongfang Zhuang,Huiyuan Chen,Liang Wang,Wei Zhang,Jeff Phillips

We propose a new mechanism to augment a word vector embedding representation that offers improved bias removal while retaining the key information—resulting in improved interpretability of the representation. Rather than removing the information associated with a concept that may induce bias, our proposed method identifies two concept subspaces and makes them orthogonal. The resulting representation has these two concepts uncorrelated. Moreover, because they are orthogonal, one can simply apply a rotation on the basis of the representation so that the resulting subspace corresponds with coordinates. This explicit encoding of concepts to coordinates works because they have been made fully orthogonal, which previous approaches do not achieve. Furthermore, we show that this can be extended to multiple subspaces. As a result, one can choose a subset of concepts to be represented transparently and explicitly, while the others are retained in the mixed but extremely expressive format of the representation.

****

Unpacking Large Language Models with Conceptual Consistency

Pritish Sahu,Michael Cogswell,Yunye Gong,Ajay Divakaran

If a Large Language Model (LLM) answers “yes” to the question “Are mountains tall?” then does it know what a mountain is? Can you rely on it responding correctly or incorrectly to other questions about mountains? The success of Large Language Models (LLMs) indicates they are increasingly able to answer queries like these accurately, but that ability does not necessarily imply a general

understanding of concepts relevant to the anchor query. We propose conceptual consistency to measure a LLM’s understanding of relevant concepts. This novel metric measures how well a model can be characterized by finding out how consistent its responses to queries about conceptually relevant background knowledge

are. To compute it we extract background knowledge by traversing paths between concepts in a knowledge base and then try to predict the model’s response to the anchor query from the background knowledge. We investigate the performance of current LLMs in a commonsense reasoning setting using the CSQA dataset and the ConceptNet knowledge base. While conceptual consistency, like other metrics, does increase with the scale of the LLM used, we find that popular models do not necessarily have high conceptual consistency. Our analysis also shows significant variation in conceptual consistency across different kinds of relations, concepts, and prompts. This serves as a step toward building models that humans can apply a theory of mind to, and thus interact with intuitively.

****

LSTM-BASED-AUTO-BI-LSTM for Remaining Useful Life (RUL) Prediction: the first round of test results

Carlos Ferreira,Gil Gonçalves

The Remaining Useful Life (RUL) is one of the most critical indicators to detect a component’s failure before it effectively occurs. It can be predicted by historical data or direct data extraction by adopting model-based, data-driven, or hybrid methodologies. Data-driven methods have mainly used Machine Learning (ML) approaches, despite several studies still pointing out different challenges in this sense. For instance, traditional ML methods cannot extract features directly from time series depending, in some cases, on the prior knowledge of the system. In this context, this work proposes a DL-based approach called LSTM-based-AUTO-Bi-LSTM. It ensembles an LSTM-based autoencoder to automatically perform feature engineering (instead of manually) with Bidirectional Long Short-Term Memory (Bi-LSTM) to predict RUL. We have tested the model using the Turbofan Engine Degradation Simulation Dataset (FD001), an open dataset. It was generated from the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) from the Prognostics Center of Excellence (PcoE), from the National Aeronautics and Space Administration (NASA). The objective is to release the first round of analytical results and statistical visualisations of the model application, which will guide us in future improvements.

****

Recurrent Real-valued Neural Autoregressive Density Estimator for Online Density Estimation and Classification of Streaming Data

Tianyu Li,Bogdan Mazoure,Guillaume Rabusseau

In contrast with the traditional offline learning, where complete data accessibility is assumed, many modern applications involve processing data in a streaming fashion. This online learning setting raises various challenges, including concept drift, hardware memory constraints, etc. In this paper, we propose the Recurrent Real-valued Neural Autoregressive Density Estimator (RRNADE), a flexible density-based model for online classification and density estimation. RRNADE combines a neural Gaussian mixture density module with a recurrent module. This combination allows RRNADE to exploit possible sequential correlations in the streaming task, which are often ignored in the classical streaming setting where each input is assumed to be independent from the previous ones. We showcase the ability of RRNADE to adapt to concept drifts on synthetic density estimation tasks. We also apply RRNADE to online classification tasks on both real world and synthetic datasets and compare it with multiple density based as well as nondensity based online classification methods. In almost all of these tasks, RRNADE outperforms the other methods. Lastly, we conduct an ablation study demonstrating the complementary benefits of the density and the recurrent modules.

****

Out-of-Distribution Detection and Selective Generation for Conditional Language Models

Jie Ren,Jiaming Luo,Yao Zhao,Kundan Krishna,Mohammad Saleh,Balaji Lakshminarayanan,Peter J Liu

Machine learning algorithms typically assume independent and identically distributed samples in training and at test time (IID). Much work has shown that high-performing ML classifiers can degrade significantly and provide overly-confident, wrong classification predictions, particularly for out-of-distribution (OOD) inputs. Conditional language models (CLMs) are predominantly trained to classify the next token in an output sequence, and may suffer even worse degradation on OOD inputs as the prediction is done auto-regressively over many steps. Furthermore, the space of potential low-quality outputs is larger as arbitrary text can be generated and it is important to know when to trust the generated output. We present a highly accurate and lightweight OOD detection method for CLMs, and demonstrate its effectiveness on abstractive summarization and translation. We also show how our method can be used under the common and realistic setting of distri bution shift for selective generation (analogous to selective prediction for classification) of high-quality outputs, while automatically abstaining from low-quality ones, enabling safer deployment of generative language models.

****

Layer Grafted Pre-training: Bridging Contrastive Learning And Masked Image Modeling For Label-Efficient Representations

Ziyu Jiang,Yinpeng Chen,Mengchen Liu,Dongdong Chen,Xiyang Dai,Lu Yuan,Zicheng Liu,Zhangyang Wang

Recently, both Contrastive Learning (CL) and Mask Image Modeling (MIM) demonstrate that self-supervision is powerful to learn good representations. However, naively combining them is far from success. In this paper, we start by making the empirical observation that a naive joint optimization of CL and MIM losses leads to conflicting gradient directions - more severe as the layers go deeper. This motivates us to shift the paradigm from combining loss at the end, to choosing the proper learning method per network layer. Inspired by experimental observations, we find that MIM and CL are suitable to lower and higher layers, respectively. We hence propose to combine them in a surprisingly simple, ``sequential cascade'' fashion: early layers are first trained under one MIM loss, on top of which latter layers continue to be trained under another CL loss. The proposed Layer Grafted Pre-training learns good visual representations that demonstrate superior label efficiency in downstream applications, in particular yielding strong few-shot performance besides linear evaluation. For instance, on ImageNet-1k, Layer Grafted Pre-training yields 65.5% Top-1 accuracy in terms of 1% few-shot learning with ViT-B/16, which improves MIM and CL baselines by 14.4% and 2.1% with no bells and whistles. The code is available at https://github.com/VITA-Group/layerGraftedPretraining_ICLR23.git.

****

Structural Adversarial Objectives for Self-Supervised Representation Learning

Xiao Zhang,Michael Maire

Within the framework of generative adversarial networks (GANs), we propose objectives that task the discriminator with additional structural modeling responsibilities. In combination with an efficient smoothness regularizer imposed on the network, these objectives guide the discriminator to learn to extract informative representations, while maintaining a generator capable of sampling from the domain. Specifically, we influence the features produced by the discriminator at two levels of granularity. At coarse scale, we impose a Gaussian assumption encouraging smoothness and diversified representation, while at finer scale, we group features forming local clusters. Experiments demonstrate that augmenting GANs with these self-supervised objectives suffices to produce discriminators which, evaluated in terms of representation learning, compete with networks trained by state-of-the-art contrastive approaches. Furthermore, operating within the GAN framework frees our system from the reliance on data augmentation schemes that is prevalent across purely contrastive representation learning methods.

****

VIMA: General Robot Manipulation with Multimodal Prompts

Yunfan Jiang,Agrim Gupta,Zichen Zhang,Guanzhi Wang,Yongqiang Dou,Yanjun Chen,Li Fei-Fei,Anima Anandkumar,Yuke Zhu,Linxi Fan

Prompt-based learning has emerged as a successful paradigm in natural language processing, where a single general-purpose language model can be instructed to perform any task specified by input prompts. Yet task specification in robotics comes in various forms, such as imitating one-shot demonstrations, following language instructions, and reaching visual goals. They are often considered different tasks and tackled by specialized models. This work shows that we can express a wide spectrum of robot manipulation tasks with *multimodal prompts*, interleaving textual and visual tokens. We design a transformer-based generalist robot agent, VIMA, that processes these prompts and outputs motor actions autoregressively. To train and evaluate VIMA, we develop a new simulation benchmark with thousands of procedurally-generated tabletop tasks with multimodal prompts, 600K+ expert trajectories for imitation learning, and four levels of evaluation protocol fo r systematic generalization. VIMA achieves strong scalability in both model capacity and data size. It outperforms prior SOTA methods in the hardest zero-shot generalization setting by up to 2.9x task success rate given the same training data. With 10x less training data, VIMA still performs 2.7x better than the top competing approach. Video demos are available at https://iclr3081.github.io/.

****

Discovering Latent Knowledge in Language Models Without Supervision

Collin Burns,Haotian Ye,Dan Klein,Jacob Steinhardt

Existing techniques for training language models can be misaligned with the truth: if we train models with imitation learning, they may reproduce errors that humans make; if we train them to generate text that humans rate highly, they may output errors that human evaluators can't detect. We propose circumventing this issue by directly finding latent knowledge inside the internal activations of a language model in a purely unsupervised way. Specifically, we introduce a method for accurately answering yes-no questions given only unlabeled model activations. It works by finding a direction in activation space that satisfies logical consistency properties, such as that a statement and its negation have opposite truth values. We show that despite using no supervision and no model outputs, our method can recover diverse knowledge represented in large language models: across 6 models and 10 question-answering datasets, it outperforms zero-shot accuracy by 4\% on average. We also find that it cuts prompt sensitivity in half and continues to maintain high accuracy even when models are prompted to generate incorrect answers. Our results provide an initial step toward discovering what language models know, distinct from what they say, even when we don't have access to explicit ground truth labels.

****

ModReduce: A Multi-Knowledge Distillation Framework with Online Learning

Yahya Saad Abbas,Abdelhakim Walid Badawy,Mohamed Mahfouz Shahawy,Samah Hussien,Samah Ayman,Hesham Mohamed Eraqi,Cherif Salama

Deep neural networks have produced revolutionary results in many applications; however, the computational resources required to use such models are expensive in terms of processing power and memory space. Research has been conducted in the field of knowledge distillation, aiming to enhance the performance of smaller models. Knowledge distillation transfers knowledge from large networks into smaller ones. Literature defines three types of knowledge that can be transferred: response-based, relational-based, and feature-based.

To the best of our knowledge, only transferring one or two types of knowledge has been studied before, but transferring all three remains unexplored. In this paper, we propose ModReduce, a framework designed to transfer the three knowledge types in a unified manner using a combination of offline and online knowledge distillation. Moreover, an extensive experimental study on the effects of combining different knowledge types on student models’ generalization and overall performance has been performed. Our experiments showed that ModReduce outperforms state-of-the-art knowledge distillation methods in terms of Average Relative Improvement.

****

Defending against Reconstruction attacks using Rényi Differential Privacy

Pierre Stock,Igor Shilov,Ilya Mironov,Alexandre Sablayrolles

Reconstruction attacks allow an adversary to regenerate data samples of the training set using access to only a trained model. It has been recently shown that simple heuristics can reconstruct data samples from language models, making this threat scenario an important aspect of model release. Differential privacy is a known solution to such attacks, but is often used with a large privacy budget (epsilon > 8) which does not translate to meaningful guarantees. In this paper we show that, for a same mechanism, we can derive privacy guarantees for reconstruction attacks that are better than the traditional ones from the literature. In particular, we show that larger privacy budgets do not provably protect against membership inference, but can still protect extraction of rare secrets. We design a method to efficiently run reconstruction attacks with lazy sampling and empirically show that we can surface at-risk training samples from non-private langua

ge models. We show experimentally that our guarantees hold on real-life language models trained with differential privacy for difficult scenarios, including GPT-2 finetuned on Wikitext-103.
****

Diffusion Adversarial Representation Learning for Self-supervised Vessel Segmentation
Boah Kim,Yujin Oh,Jong Chul Ye
Vessel segmentation in medical images is one of the important tasks in the diagnosis of vascular diseases and therapy planning. Although learning-based segmentation approaches have been extensively studied, a large amount of ground-truth labels are required in supervised methods and confusing background structures make neural networks hard to segment vessels in an unsupervised manner. To address this, here we introduce a novel diffusion adversarial representation learning (DARL) model that leverages a denoising diffusion probabilistic model with adversarial learning, and apply it to vessel segmentation. In particular, for self-supervised vessel segmentation, DARL learns the background signal using a diffusion module, which lets a generation module effectively provide vessel representations. Also, by adversarial learning based on the proposed switchable spatially-adaptive denormalization, our model estimates synthetic fake vessel images as well as vessel segmentation masks, which further makes the model capture vessel-relevant semantic information. Once the proposed model is trained, the model generates segmentation masks in a single step and can be applied to general vascular structure segmentation of coronary angiography and retinal images. Experimental results on various datasets show that our method significantly outperforms existing unsupervised and self-supervised vessel segmentation methods.
****

Reconciling Security and Communication Efficiency in Federated Learning
Karthik Prasad,Sayan Ghosh,Graham Cormode,Ilya Mironov,Ashkan Yousefpour,Pierre Stock
Cross-device Federated Learning is an increasingly popular machine learning setting to train a model by leveraging a large population of client devices with high privacy and security guarantees. However, communication efficiency remains a major bottleneck when scaling federated learning to production environments, particularly due to bandwidth constraints during uplink communication. In this paper, we formalize and address the problem of compressing client-to-server model updates under the Secure Aggregation primitive, a core component of Federated Learning pipelines that allows the server to aggregate the client updates without accessing them individually. In particular, we adapt standard scalar quantization and pruning methods to Secure Aggregation and propose Secure Indexing, a variant of Secure Aggregation that supports quantization for extreme compression. We establish state-of-the-art results on LEAF benchmarks in a secure Federated Learning setup with up to 40x compression in uplink communication with no meaningful loss in utility compared to uncompressed baselines.
****

Semantic Image Manipulation with Background-guided Internal Learning
Zhongping Zhang,Huiwen He,Bryan A. Plummer,Zhenyu Liao,Huayan Wang
Image manipulation has attracted a lot of interest due to its wide range of applications. Prior work modifies images either from low-level manipulation, such as image inpainting or through manual edits via paintbrushes and scribbles, or from high-level manipulation, employing deep generative networks to output an image conditioned on high-level semantic input. In this study, we propose Semantic Image Manipulation with Background-guided Internal Learning (SIMBIL), which combines high-level and low-level manipulation. Specifically, users can edit an image at the semantic level by applying changes on the scene graph. Then our model manipulates the image at the pixel level according to the modified scene graph. There are two major advantages of our approach. First, high-level manipulation requires less manual effort from the user compared to manipulating raw image pixels. Second, our low-level internal learning approach is scalable to images of various sizes without reliance on external visual datasets for training. We outperform the state-of-the-art in a quantitative and qualitative evaluation on CLEVR and

Visual Genome datasets. Experiments show around 8 points improvement of SSIM (RoI) on CLEVR and around 25% improvement of user evaluation accuracy on Visual Genome, demonstrating the effectiveness of our approach.

****

Pretraining the Vision Transformer using self-supervised methods for vision based Deep Reinforcement Learning

Manuel Goulão,Arlindo L. Oliveira

The Vision Transformer architecture has shown to be competitive in the computer vision (CV) space where it has dethroned convolution-based networks in several benchmarks. Nevertheless, Convolutional Neural Networks (CNN) remain the preferential architecture for the representation module in Reinforcement Learning. In this work, we study pretraining a Vision Transformer using several state-of-the-art self-supervised methods and assess data-efficiency gains from this training framework. We propose a new self-supervised learning method called TOV-VICReg that extends VICReg to better capture temporal relations between observations by adding a temporal order verification task. Furthermore, we evaluate the resultant encoders with Atari games in a sample-efficiency regime. Our results show that the vision transformer, when pretrained with TOV-VICReg, outperforms the other self-supervised methods but still struggles to overcome a CNN. Nevertheless, we were able to outperform a CNN in two of the ten games where we perform a 100k steps evaluation. Ultimately, we believe that such approaches in Deep Reinforcement Learning (DRL) might be the key to achieving new levels of performance as seen in natural language processing and computer vision.

****

Noise Injection Node Regularization for Robust Learning

Noam Itzhak Levi,Itay Mimouni Bloch,Marat Freytsis,Tomer Volansky

We introduce Noise Injection Node Regularization (NINR), a method of injecting structured noise into Deep Neural Networks (DNN) during the training stage, resulting in an emergent regularizing effect. We present theoretical and empirical evidence for substantial improvement in robustness against various test data perturbations for feed-forward DNNs when trained under NINR. The novelty in our approach comes from the interplay of adaptive noise injection and initialization conditions such that noise is the dominant driver of dynamics at the start of training. As it simply requires the addition of external nodes without altering the existing network structure or optimization algorithms, this method can be easily incorporated into many standard problem specifications. We find improved stability against a number of data perturbations, including domain shifts, with the most dramatic improvement obtained for unstructured noise, where our technique outperforms other existing methods such as dropout or L_2 regularization, in some cases. We further show that desirable generalization properties on clean data are generally maintained.

****

Q-Ensemble for Offline RL: Don't Scale the Ensemble, Scale the Batch Size

Alexander Nikulin,Vladislav Kurenkov,Denis Tarasov,Dmitry Akimov,Sergey Kolesnikov

Training large neural networks is known to be time-consuming, where learning duration may stretch up to days or weeks. To address this problem, the approach of large-batch optimization was introduced, demonstrating that scaling mini-batch sizes with appropriate learning rate adjustments may speed up the training process by orders of magnitude. While long training time was not typically a major issue for model-free deep offline RL algorithms, recently introduced Q-ensemble methods achieving state-of-the-art performance made this issue more relevant, notably extending the training duration. In this work, we demonstrate how large-batch optimization, typically overlooked in deep offline RL community, can benefit this class of methods. We show that simply scaling the mini-batch size and naively adjusting the learning rate allows for (1) a reduced size of the Q-ensemble, (2) stronger penalization of out-of-distribution actions, and (3) improved convergence time, effectively shortening training durations by 2.5x times on average.

****

Efficient Edge Inference by Selective Query

Anil Kag,Igor Fedorov,Aditya Gangrade,Paul Whatmough,Venkatesh Saligrama

Edge devices provide inference on predictive tasks to many end-users. However, deploying deep neural networks that achieve state-of-the-art accuracy on these devices is infeasible due to edge resource constraints. Nevertheless, cloud-only processing, the de-facto standard, is also problematic, since uploading large amounts of data imposes severe communication bottlenecks. We propose a novel end-to-end hybrid learning framework that allows the edge to selectively query only those hard examples that the cloud can classify correctly. Our framework optimizes over neural architectures and trains edge predictors and routing models so that the overall accuracy remains high while minimizing the overall latency. Training a hybrid learner is difficult since we lack annotations of hard edge-examples. We introduce a novel proxy supervision in this context and show that our method adapts seamlessly and near optimally across different latency regimes. On the ImageNet dataset, our proposed method deployed on a micro-controller unit exhibits 25% reduction in latency compared to cloud-only processing while suffering no excess loss.

****

Learning Intuitive Policies Using Action Features

Mingwei Ma,Jizhou Liu,Samuel Sokota,Max Kleiman-Weiner,Jakob Nicolaus Foerster

An unaddressed challenge in multi-agent coordination is to enable AI agents to exploit the semantic relationships between the features of actions and the features of observations. Humans take advantage of these relationships in highly intuitive ways. For instance, in the absence of a shared language, we might point to the object we desire or hold up our fingers to indicate how many objects we want. To address this challenge, we investigate the effect of network architecture on the propensity of learning algorithms to exploit these semantic relationships. Across a procedurally generated coordination task, we find that attention-based architectures that jointly process a featurized representation of observations and actions have a better inductive bias for zero-shot coordination. Through fine-grained evaluation and scenario analysis, we show that the resulting policies are human-interpretable. Moreover, such agents coordinate with people without training on any human data.

****

Differentially Private L_2-Heavy Hitters in the Sliding Window Model

Jeremiah Blocki,Seunghoon Lee,Tamalika Mukherjee,Samson Zhou

The data management of large companies often prioritize more recent data, as a source of higher accuracy prediction than outdated data. For example, the Facebook data policy retains user search histories for 6 months while the Google data retention policy states that browser information may be stored for up to 9 months. These policies are captured by the sliding window model, in which only the most recent W statistics form the underlying dataset. In this paper, we consider the problem of privately releasing the L_2-heavy hitters in the sliding window model, which include L_p-heavy hitters for $p\le 2$ and in some sense are the strongest possible guarantees that can be achieved using polylogarithmic space, but cannot be handled by existing techniques due to the sub-additivity of the L_2 norm. Moreover, existing non-private sliding window algorithms use the smooth histogram framework, which has high sensitivity. To overcome these barriers, we introduce the first differentially private algorithm for L_2-heavy hitters in the sliding window model by initiating a number of L_2-heavy hitter algorithms across the stream with significantly lower threshold. Similarly, we augment the algorithms with an approximate frequency tracking algorithm with significantly higher accuracy. We then use smooth sensitivity and statistical distance arguments to show that we can add noise proportional to an estimation of the L_2 norm. To the best of our knowledge, our techniques are the first to privately release statistics that are related to a sub-additive function in the sliding window model, and may be of independent interest to future differentially private algorithmic design in the sliding window model.

****

Scaling Convex Neural Networks with Burer-Monteiro Factorization

Arda Sahiner,Tolga Ergen,Batu Ozturkler,John M. Pauly,Morteza Mardani,Mert Pilan

ci
Recently, it has been demonstrated that a wide 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, thereby 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 demonstrate the utility and implications of the novel relative optimality bound for stationary points of the Burer-Monteiro factorization.
****

Human-level Atari 200x faster
Steven Kapturowski,Víctor Campos,Ray Jiang,Nemanja Rakicevic,Hado van Hasselt,Charles Blundell,Adria Puigdomenech Badia
The task of building general agents that perform well over a wide range of tasks has been an important goal in reinforcement learning since its inception. The problem has been subject of research of a large body of work, with performance frequently measured by observing scores over the wide range of environments contained in the Atari 57 benchmark. Agent57 was the first agent to surpass the human benchmark on all 57 games, but this came at the cost of poor data-efficiency, requiring nearly 80 billion frames of experience to achieve. Taking Agent57 as a starting point, we employ a diverse set of strategies to achieve a 200-fold reduction of experience needed to outperform the human baseline, within our novel agent MEME. We investigate a range of instabilities and bottlenecks we encountered while reducing the data regime, and propose effective solutions to build a more robust and efficient agent. We also demonstrate competitive performance with high-performing methods such as Muesli and MuZero. Our contributions aim to achieve faster propagation of learning signals related to rare events, stabilize learning under differing value scales, improve the neural network architecture, and make updates more robust under a rapidly-changing policy.
****

Wide Graph Neural Network
Jiaqi Sun,Lin Zhang,Guangyi Chen,Kun Zhang,Peng XU,Yujiu Yang
Usually, graph neural networks (GNNs) suffer from several problems, e.g., over-smoothing (in the spatial domain), poor flexibility (in the spectral domain), and low performance on heterophily (in both domains). In this paper, we provide a new GNN framework, called Wide Graph Neural Networks (WGNN) to solve these problems. It is motivated by our proposed unified view of GNNs from the perspective of dictionary learning. In light of this view, we formulate the graph learning in GNNs as learning representations from the dictionaries, where the fixed graph information is regarded as the dictionary and the trainable parameters are representations. Then, the dictionaries of spatial GNNs encode the adjacency matrix multiplication, while spectral ones sum its polynomials. Differently, WGNN directly concatenates all polynomials as the dictionary, where each polynomial is a sub-dictionary. Beyond polynomials, WGNN allows sub-dictionaries with an arbitrary size, for instance, the principal components of the adjacency matrix. This wide concatenation structure enjoys the great capability of avoiding over-smoothing and promoting flexibility, while the supplement of principal components can significantly improve the representation of heterophilic graphs. We provide a detailed theoretical analysis and conduct extensive experiments on eight datasets to demonstrate the superiority of the proposed WGNN.
****

Taming the Long Tail of Deep Probabilistic Forecasting

Mayank Sharan,Jedrzej Kozerawski,Rose Yu

Deep probabilistic forecasting is gaining attention in numerous applications from weather prognosis, through electricity consumption estimation, to autonomous vehicle trajectory prediction. However, existing approaches focus on improvements on average metrics without addressing the long tailed distribution of errors. In this work, we observe long tail behavior in the error distribution of state-of-the-art deep learning methods for probabilistic forecasting. We present two loss augmentation methods to reduce tailedness: Pareto Loss and Kurtosis Loss. Both methods are related to the concept of moments, which measures the shape of a distribution. Kurtosis Loss is based on a symmetric measure, the fourth moment. Pareto Loss is based on an asymmetric measure of right tailedness and models loss using a Generalized Pareto Distribution (GPD). We demonstrate the performance of our methods on several real-world datasets, including time series and spatiotemporal trajectories, achieving significant improvements on tail error metrics, while maintaining and even improving upon average error metrics.

****

Approximate Conditional Coverage via Neural Model Approximations

Allen Schmaltz,Danielle Rasooly

We propose a new approach for constructing prediction sets for Transformer networks via the strong signals for prediction reliability from KNN-based approximations. This enables a data-driven partitioning of the high-dimensional feature space and a new Inductive Venn Predictor for calibration, the Venn-ADMIT Predictor. Our approach more closely obtains approximate conditional coverage than recent work proposing adaptive and localized conformal score functions for deep networks. We analyze coverage on several representative natural language processing classification tasks, including class-imbalanced and distribution-shifted settings.

****

AUTOJOIN: EFFICIENT ADVERSARIAL TRAINING FOR ROBUST MANEUVERING VIA DENOISING AUTOEN- CODER AND JOINT LEARNING

Taylor Michael Villarreal,Bibek Poudel,Ryan Wickman,Yu Shen,Weizi Li

As a result of increasingly adopted machine learning algorithms and ubiquitous sensors, many ‘perception-to-control’ systems are developed and deployed. For these systems to be trustworthy, we need to improve their robustness with adversarial training being one approach. We propose a gradient-free adversarial training technique, called AutoJoin, which is a very simple yet effective and efficient approach to produce robust models for imaged-based maneuvering. Compared to other SOTA methods with testing on over 5M perturbed and clean images, AutoJoin achieves significant performance increases up to the 40% range under gradient-free perturbations while improving on clean performance up to 300%. Regarding efficiency, AutoJoin demonstrates strong advantages over other SOTA techniques by saving up to 83% time per training epoch and 90% training data. Although not the focus of AutoJoin, it even demonstrates superb ability in defending gradient-based attacks. The core idea of AutoJoin is to use a decoder attachment to the original regression model creating a denoising autoencoder within the architecture. This architecture allows the tasks ‘maneuvering’ and ‘denoising sensor input’ to be jointly learnt and reinforce each other’s performance.

****

Private Data Stream Analysis for Universal Symmetric Norm Estimation

Vladimir Braverman,Joel Manning,Steven Wu,Samson Zhou

We study how to release summary statistics on a data stream subject to the constraint of differential privacy. In particular, we focus on releasing the family of \emph{symmetric norms}, which are invariant under sign-flips and coordinate-wise permutations on an input data stream and include L_p norms, k-support norms, top-k norms, and the box norm as special cases. Although it may be possible to design and analyze a separate mechanism for each symmetric norm, we propose a general parametrizable framework that differentially privately releases a number of sufficient statistics from which the approximation of all symmetric norms can be simultaneously computed. Our framework partitions the coordinates of the underlying frequency vector into different levels based on their magnitude and releases approximate frequencies for the ``heavy'' coordinates in important leve

ls and releases approximate level sizes for the ``light'' coordinates in important levels. Surprisingly, our mechanism allows for the release of an \emph{arbitrary} number of symmetric norm approximations without any overhead or additional loss in privacy. Moreover, our mechanism permits $(1+\alpha)$-approximation to each of the symmetric norms and can be implemented using sublinear space in the streaming model for many regimes of the accuracy and privacy parameters.
****
StarGraph: Knowledge Representation Learning based on Incomplete Two-hop Subgraph
Hongzhu Li,Xiangrui Gao,Linhui Feng,Yafeng Deng,Yuhui Yin
Conventional representation learning algorithms for knowledge graphs (KG) map each entity to a unique embedding vector, ignoring the rich information contained in the neighborhood. We propose a method named StarGraph, which gives a novel way to utilize the neighborhood information for large-scale knowledge graphs to obtain entity representations. An incomplete two-hop neighborhood subgraph for each target node is at first generated, then processed by a modified self-attention network to obtain the entity representation, which is used to replace the entity embedding in conventional methods. We achieved SOTA performance on ogbl-wikikg2 and got competitive results on fb15k-237. The experimental results proves that StarGraph is efficient in parameters, and the improvement made on ogbl-wikikg2 demonstrates its great effectiveness of representation learning on large-scale knowledge graphs.
****
Temporal Domain Generalization with Drift-Aware Dynamic Neural Networks
Guangji Bai,Chen Ling,Liang Zhao
Temporal domain generalization is a promising yet extremely challenging area where the goal is to learn models under temporally changing data distributions and generalize to unseen data distributions following the trends of the change. The advancement of this area is challenged by: 1) characterizing data distribution drift and its impacts on models, 2) expressiveness in tracking the model dynamics, and 3) theoretical guarantee on the performance. To address them, we propose a Temporal Domain Generalization with Drift-Aware Dynamic Neural Network (DRAIN) framework. Specifically, we formulate the problem into a Bayesian framework that jointly models the relation between data and model dynamics. We then build a recurrent graph generation scenario to characterize the dynamic graph-structured neural networks learned across different time points. It captures the temporal drift of model parameters and data distributions and can predict models in the future without the presence of future data. In addition, we explore theoretical guarantees of the model performance under the challenging temporal DG setting and provide theoretical analysis, including uncertainty and generalization error. Finally, extensive experiments on several real-world benchmarks with temporal drift demonstrate the proposed method’s effectiveness and efficiency.
****
Incompatibility Clustering as a Defense Against Backdoor Poisoning Attacks
Charles Jin,Melinda Sun,Martin Rinard
We propose a novel clustering mechanism based on an incompatibility property between subsets of data that emerges during model training. This mechanism partitions the dataset into subsets that generalize only to themselves, i.e., training on one subset does not improve performance on the other subsets. Leveraging the interaction between the dataset and the training process, our clustering mechanism partitions datasets into clusters that are defined by—and therefore meaningful to—the objective of the training process.

We apply our clustering mechanism to defend against data poisoning attacks, in which the attacker injects malicious poisoned data into the training dataset to affect the trained model's output. Our evaluation focuses on backdoor attacks against deep neural networks trained to perform image classification using the GTSRB and CIFAR-10 datasets. Our results show that (1) these attacks produce poisoned datasets in which the poisoned and clean data are incompatible and (2) our technique successfully identifies (and removes) the poisoned data. In an end-to-end evaluation, our defense reduces the attack success rate to below 1% on 134 out of 165 scenarios, with only a 2% drop in clean accuracy on CIFAR-10 and a negligible drop in clean accuracy on GTSRB.

****

Enforcing Hard Constraints with Soft Barriers: Safe Reinforcement Learning in Unknown Stochastic Environments

Yixuan Wang,Simon Sinong Zhan,Ruochen Jiao,Zhilu Wang,Wanxin Jin,Zhuoran Yang,Zhaoran Wang,Chao Huang,Qi Zhu

It is quite challenging to ensure the safety of reinforcement learning (RL) agents in an unknown and stochastic environment under hard constraints that require the system state not to reach certain specified unsafe regions. Many popular safe RL methods such as those based on the Constrained Markov Decision Process (CMDP) paradigm formulate safety violations in a cost function and try to constrain the expectation of cumulative cost under a threshold. However, it is often difficult to effectively capture and enforce hard reachability-based safety constraints indirectly with such constraints on safety violation cost. In this work, we leverage the notion of barrier function to explicitly encode the hard safety constraints, and given that the environment is unknown, relax them to our design of \emph{generative-model-based soft barrier functions}. Based on such soft barriers, we propose a safe RL approach that can jointly learn the environment and optimize the control policy, while effectively avoiding the unsafe regions with safety probability optimization.

Experiments on a set of examples demonstrate that our approach can effectively enforce hard safety constraints and significantly outperform CMDP-based baseline methods in system safe rate measured via simulations.

****

Integrating Episodic and Global Novelty Bonuses for Efficient Exploration

Mikael Henaff,Minqi Jiang,Roberta Raileanu

Exploration in environments which differ across episodes has received increasing attention in recent years. Current methods use some combination of global novelty bonuses, computed using the agent's entire training experience, and episodic novelty bonuses, computed using only experience from the current episode. However, the use of these two types of bonuses has been ad-hoc and poorly understood. In this work, we first shed light on the behavior these two kinds of bonuses on hard exploration tasks through easily interpretable examples. We find that the two types of bonuses succeed in different settings, with episodic bonuses being most effective when there is little shared structure between environments and global bonuses being effective when more structure is shared. We also find that combining the two bonuses leads to more robust behavior across both of these settings. Motivated by these findings, we then investigate different algorithmic choices for defining and combining function approximation-based global and episodic bonuses. This results in a new algorithm which sets a new state of the art across 18 tasks from the MiniHack suite used in prior work. Our code is public at \url{web-link}.

****

A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games

Samuel Sokota,Ryan D'Orazio,J Zico Kolter,Nicolas Loizou,Marc Lanctot,Ioannis Mitliagkas,Noam Brown,Christian Kroer

This work studies an algorithm, which we call magnetic mirror descent, that is inspired by mirror descent and the non-Euclidean proximal gradient algorithm. Our contribution is demonstrating the virtues of magnetic mirror descent as both an equilibrium solver and as an approach to reinforcement learning in two-player zero-sum games. These virtues include: 1) Being the first quantal response equilibria solver to achieve linear convergence for extensive-form games with first order feedback; 2) Being the first standard reinforcement learning algorithm to achieve empirically competitive results with CFR in tabular settings; 3) Achieving favorable performance in 3x3 Dark Hex and Phantom Tic-Tac-Toe as a self-play deep reinforcement learning algorithm.

****

Dynamics-aware Skill Generation from Behaviourally Diverse Demonstrations
Shibei Zhu,Rituraj Kaushik,Samuel Kaski,Ville Kyrki
Learning from demonstrations (LfD) provides a data-efficient way for a robot to learn a task by observing humans performing the task, without the need for an explicit reward function. However, in many real-world scenarios (e.g., driving a car) humans often perform the same task in different ways, motivated not only by the primary objective of the task (e.g., reaching the destination safely) but also by their individual preferences (e.g., different driving behaviours), leading to a multi-modal distribution of demonstrations. In this work, we consider an LfD problem, where the reward function for the main objective of the task is known to the learning agent; however, the individual preferences leading to the variations in the demonstrations are unknown. We show that current LfD approaches learn policies that either track a single mode or the mean of the demonstration distribution. In contrast, we propose an algorithm to learn a diverse set of policies to perform the task, capturing the different modes in the demonstrations due to the diverse preferences of the individuals. We show that we can build a parameterised solution space that captures different behaviour patterns from the demonstrations. Then, a set of policies can be generated in solution space that generate a diverse range of behaviours that go beyond the provided demonstrations.
****

DiP-GNN: Discriminative Pre-Training of Graph Neural Networks
Simiao Zuo,Haoming Jiang,Qingyu Yin,Xianfeng Tang,Bing Yin,Tuo Zhao
Graph neural network (GNN) pre-training methods have been proposed to enhance the power of GNNs. Specifically, a GNN is first pre-trained on a large-scale unlabeled graph and then fine-tuned on a separate small labeled graph for downstream applications, such as node classification. One popular pre-training method is to mask out a proportion of the edges, and a GNN is trained to recover them. However, such a generative method suffers from graph mismatch. That is, the masked graph input to the GNN deviates from the original graph. To alleviate this issue, we propose DiP-GNN (Discriminative Pre-training of Graph Neural Networks). Specifically, we train a generator to recover identities of the masked edges, and simultaneously, we train a discriminator to distinguish the generated edges from the original graph's edges. The discriminator is subsequently used for downstream fine-tuning. In our pre-training framework, the graph seen by the discriminator better matches the original graph because the generator can recover a proportion of the masked edges. Extensive experiments on large-scale homogeneous and heterogeneous graphs demonstrate the effectiveness of the proposed framework. Our code will be publicly available.
****

Learning to Act through Activation Function Optimization in Random Networks
Joachim Winther Pedersen,Sebastian Risi
Biological neural networks are characterised by a high degree of neural diversity, a trait that artificial neural networks (ANNs) generally lack.
Additionally, learning in ANNs is typically synonymous with only modifying the strengths of connection weights.
However, there is much evidence from neuroscience that different classes of neurons each have crucial roles in the information processing done by the network. In nature, each neuron is a dynamical system that is a powerful information processor in its own right. In this paper we ask the question, how well can ANNs learn to perform reinforcement learning tasks only through the optimization of neural activation functions, without any weight optimization?
We demonstrate the viability of the method and show that the neural parameters are expressive enough to allow learning three different continuous control tasks without weight optimization.
These results open up for more possibilities for synergies between synaptic and neural optimization in ANNs in the future.
****

Safer Reinforcement Learning with Counterexample-guided Offline Training
Xiaotong Ji,Antonio Filieri
Safe reinforcement learning (RL) aims at addressing the limitation of reinforcem

ent learning in safety-critical scenarios, where failures during learning may incur high costs. Several methods exist to incorporate external knowledge or to use proximal sensor data to limit the exploration of unsafe states. However, dealing with (partially) unknown environments and dynamics, where an agent must discover safety threats during exploration, remains challenging. In this paper, we propose a method to abstract hybrid continuous-discrete systems into compact surrogate models representing the safety-relevant knowledge acquired by the agent at any time during exploration. We exploit probabilistic counterexamples generation to synthesise minimal, partial simulation environments from the surrogate model where the agent can train offline to produce heuristic strategies to minimise the risk of visiting unsafe states during subsequent online exploration. We demonstrate our method's effectiveness in increasing the agent's exploration safety on a selection of OpenAI Gym benchmarks.
****
Pitfalls of Gaussians as a noise distribution in NCE
Holden Lee,Chirag Pabbaraju,Anish Prasad Sevekari,Andrej Risteski
Noise Contrastive Estimation (NCE) is a popular approach for learning probability density functions parameterized up to a constant of proportionality. The main idea is to design a classification problem for distinguishing training data from samples from an (easy-to-sample) noise distribution q, in a manner that avoids having to calculate a partition function. It is well-known that the choice of q can severely impact the computational and statistical efficiency of NCE. In practice, a common choice for q is a Gaussian which matches the mean and covariance of the data.

In this paper, we show that such a choice can result in an exponentially bad (in the ambient dimension) conditioning of the Hessian of the loss - even for very simple data distributions. As a consequence, both the statistical and algorithmic complexity for such a choice of q will be problematic in practice - suggesting that more complex noise distributions are essential to the success of NCE.
****
Scaling Laws for a Multi-Agent Reinforcement Learning Model
Oren Neumann,Claudius Gros
The recent observation of neural power-law scaling relations has made a significant impact in the field of deep learning. A substantial amount of attention has been dedicated as a consequence to the description of scaling laws, although mostly for supervised learning and only to a reduced extent for reinforcement learning frameworks. In this paper we present an extensive study of performance scaling for a cornerstone reinforcement learning algorithm, AlphaZero. On the basis of a relationship between Elo rating, playing strength and power-law scaling, we train AlphaZero agents on the games Connect Four and Pentago and analyze their performance. We find that player strength scales as a power law in neural network parameter count when not bottlenecked by available compute, and as a power of compute when training optimally sized agents. We observe nearly identical scaling exponents for both games. Combining the two observed scaling laws we obtain a power law relating optimal size to compute similar to the ones observed for language models. We find that the predicted scaling of optimal neural network size fits our data for both games. This scaling law implies that previously published state-of-the-art game-playing models are significantly smaller than their optimal size, given the respective compute budgets. We also show that large AlphaZero models are more sample efficient, performing better than smaller models with the same amount of training data.
****
Risk Control for Online Learning Models
Shai Feldman,Liran Ringel,Stephen Bates,Yaniv Romano
To provide rigorous uncertainty quantification for online learning models, we develop a framework for constructing uncertainty sets that provably control risk---such as coverage of confidence intervals, false negative rate, or F1 score---in the online setting. This extends conformal prediction to apply to a larger class of online learning problems. Our method guarantees risk control at any user-sp

ecified level even when the underlying data distribution shifts drastically, even adversarially, over time in an unknown fashion.
The technique we propose is highly flexible as it can be applied with any base online learning algorithm (e.g., a deep neural network trained online), requiring minimal implementation effort and essentially zero additional computational cost.
We further extend our approach to control multiple risks simultaneously, so the prediction sets we generate are valid for all given risks.
To demonstrate the utility of our method, we conduct experiments on real-world tabular time-series data sets showing that the proposed method rigorously controls various natural risks.
Furthermore, we show how to construct valid intervals for an online image-depth estimation problem that previous sequential calibration schemes cannot handle.
****

Federated Learning with Openset Noisy Labels
Zonglin Di,Zhaowei Zhu,Xin Eric Wang,Yang Liu
Federated learning is a learning paradigm that allows the central server to learn from different data sources while keeping the data private at local. Without controlling and monitoring the local data collection process, it is highly likely that the locally available training labels are noisy, just as in a centralized data collection effort. Moreover, different clients may hold samples within different label spaces. The noisy label space is likely to be different from the unobservable clean label space, resulting in openset noisy labels. In this work, we study the challenge of federated learning from clients with openset noisy labels. We observe that many existing solutions, e.g., loss correction, in the noisy label literature cannot achieve their originally claimed effect in local training. A central contribution of this work is to propose an approach that communicates globally randomly selected ``contrastive labels" among clients to prevent local models from memorizing the openset noise patterns individually. Randomized label generations are applied during label sharing to facilitate access to the contrastive labels while ensuring differential privacy (DP). Both the DP guarantee and the effectiveness of our approach are theoretically guaranteed. Compared with several baseline methods, our solution shows its efficiency in several public benchmarks and real-world datasets under different noise ratios and noise models.
****

Perfectly Secure Steganography Using Minimum Entropy Coupling
Christian Schroeder de Witt,Samuel Sokota,J Zico Kolter,Jakob Nicolaus Foerster,Martin Strohmeier
Steganography is the practice of encoding secret information into innocuous content in such a manner that an adversarial third party would not realize that there is hidden meaning. While this problem has classically been studied in security literature, recent advances in generative models have led to a shared interest among security and machine learning researchers in developing scalable steganography techniques. In this work, we show that a steganography procedure is perfectly secure under Cachin (1998)'s information theoretic-model of steganography if and only if it is induced by a coupling. Furthermore, we show that, among perfectly secure procedures, a procedure is maximally efficient if and only if it is induced by a minimum entropy coupling. These insights yield what are, to the best of our knowledge, the first steganography algorithms to achieve perfect security guarantees with non-trivial efficiency; additionally, these algorithms are highly scalable. To provide empirical validation, we compare a minimum entropy coupling-based approach to three modern baselines---arithmetic coding, Meteor, and adaptive dynamic grouping---using GPT-2, WaveRNN, and Image Transformer as communication channels. We find that the minimum entropy coupling-based approach achieves superior encoding efficiency, despite its stronger security constraints. In aggregate, these results suggest that it may be natural to view information-theoretic steganography through the lens of minimum entropy coupling.
****

The power of choices in decision tree learning

Guy Blanc,Jane Lange,Chirag Pabbaraju,Li-Yang Tan

We propose a simple and natural generalization of standard and empirically successful decision tree learning algorithms such as ID3, C4.5, and CART.   These classic algorithms, which have been central to machine learning for decades, are greedy in nature: they grow a decision tree by iteratively splitting on the "best" attribute. We augment these algorithms with an additional greediness parameter k and our resulting algorithm, Top-k, considers the k best attributes as possible splits instead of just the single best attribute.

We demonstrate, theoretically and empirically, the power of this simple generalization. We first prove a sharp greediness hierarchy theorem showing that for every $k\in \mathbb{N}$, Top-$(k+1)$ can be much more powerful than Top-k: there are data distributions for which the former achieves accuracy $1-\epsilon$, whereas the latter only achieves accuracy $\frac{1}{2}+\epsilon$. We then show, through extensive experiments, that Top-k compares favorably with the two main approaches to decision tree learning: classic greedy algorithms and more recent "optimal decision tree" algorithms. On one hand, Top-k consistently enjoys significant accuracy gains over the greedy algorithms across a wide range of benchmarks, at the cost of only a mild training slowdown. On the other hand, Top-k is markedly more scalable than optimal decision tree algorithms, and is able to handle dataset and feature set sizes that remain beyond the reach of these algorithms.

Taken together, our results highlight the potential practical impact of the power of choices in decision tree learning.

****

Identifiability of Label Noise Transition Matrix

Yang Liu,Hao Cheng,Kun Zhang

The noise transition matrix plays a central role in the problem of learning with noisy labels. Among many other reasons, a large number of existing solutions rely on access to it. Identifying and estimating the transition matrix without ground truth labels is a critical and challenging task. When label noise transition depends on each instance, the problem of identifying the instance-dependent noise transition matrix becomes substantially more challenging. Despite recent works proposing solutions for learning from instance-dependent noisy labels, the field lacks a unified understanding of when such a problem remains identifiable. The goal of this paper is to characterize the identifiability of the label noise transition matrix. Building on Kruskal's identifiability results, we show the necessity of multiple noisy labels in identifying the noise transition matrix for the generic case at the instance level. We further instantiate the results to relate to the successes of the state-of-the-art solutions and how additional assumptions alleviated the requirement of multiple noisy labels. Our result also reveals that disentangled features are helpful in the above identification task and we provide empirical evidence.

****

Learning from Others: Similarity-based Regularization for Mitigating Artifacts

Reda Igbaria,Yonatan Belinkov

Common methods for mitigating spurious correlations in natural language understanding (NLU) usually operate in the output space, encouraging a main model to behave differently from a bias model by down-weighing examples where the bias model is confident.

While improving out of distribution (OOD) performance, it was recently observed that the internal representations of the presumably debiased models are actually more, rather than less biased.

We propose SimgReg, a new method for debiasing internal model components via similarity-based regularization, in representation space: We encourage the model to learn representations that are either similar to an unbiased model or different from a biased model. We experiment with three NLU tasks and different kinds of biases.

We find that SimReg improves OOD performance, with little in-distribution degrad ation. Moreover, the representations learned by SimReg are less biased than in other methods.

****

Calibrating Transformers via Sparse Gaussian Processes
Wenlong Chen,Yingzhen Li
Transformer models have achieved profound success in prediction tasks in a wide range of applications in natural language processing, speech recognition and computer vision. Extending Transformer’s success to safety-critical domains requires calibrated uncertainty estimation which remains under-explored. To address this, we propose Sparse Gaussian Process attention (SGPA), which performs Bayesian inference directly in the output space of multi-head attention blocks (MHAs) in transformer to calibrate its uncertainty. It replaces the scaled dot-product operation with a valid symmetric kernel and uses sparse Gaussian processes (SGP) techniques to approximate the posterior processes of MHA outputs. Empirically, on a suite of prediction tasks on text, images and graphs, SGPA-based Transformers achieve competitive predictive accuracy, while noticeably improving both in-distribution calibration and out-of-distribution robustness and detection.

****

Model Transferability with Responsive Decision Subjects
Yang Liu,Yatong Chen,Zeyu Tang,Kun Zhang
This paper studies model transferability when human decision subjects respond to a deployed machine learning model. In our setting, an agent or a user corresponds to a sample (X,Y) drawn from a distribution $\mathcal{D}$ and will face a model h and its classification result $h(X)$. Agents can modify X to adapt to h, which will incur a distribution shift on (X,Y). Therefore, when training h, the learner will need to consider the subsequently ``induced" distribution when the output model is deployed. Our formulation is motivated by applications where the deployed machine learning models interact with human agents, and will ultimately face \emph{responsive} and interactive data distributions. We formalize the discussions of the transferability of a model by studying how the model trained on the available source distribution (data) would translate to the performance on the induced domain. We provide both upper bounds for the performance gap due to the induced domain shift, as well as lower bound for the trade-offs that a classifier has to suffer on either the source training distribution or the induced target distribution. We provide further instantiated analysis for two popular domain adaptation settings with covariate shift and target shift.

****

Red PANDA: Disambiguating Image Anomaly Detection by Removing Nuisance Factors
Niv Cohen,Jonathan Kahana,Yedid Hoshen
Anomaly detection methods strive to discover patterns that differ from the norm in a meaningful way. This goal is ambiguous as different human operators may find different attributes meaningful. An image differing from the norm by an attribute such as pose may be considered anomalous by some operators while others may consider the attribute irrelevant. Breaking from previous research, we present a new anomaly detection method that allows operators to exclude an attribute when detecting anomalies. Our approach aims to learn representations which do not contain information regarding such nuisance attributes. Anomaly scoring is performed using a density-based approach. Importantly, our approach does not require specifying the attributes where anomalies could appear, which is typically impossible in anomaly detection, but only attributes to ignore. An empirical investigation is presented verifying the effectiveness of our approach.

****

Abstracting Imperfect Information Away from Two-Player Zero-Sum Games
Samuel Sokota,Ryan D'Orazio,Chun Kai Ling,David J Wu,J Zico Kolter,Noam Brown
In their seminal work, Nayyar et al. (2013) showed that imperfect information can be abstracted away from common-payoff games by having players publicly announce their policies as they play. This insight underpins sound solvers and decision-time planning algorithms for common-payoff games. Unfortunately, a naive application of the same insight to two-player zero-sum games fails because Nash equili

bria of the game with public policy announcements may not correspond to Nash equilibria of the original game. As a consequence, existing sound decision-time planning algorithms require complicated additional mechanisms that have unappealing properties. The main contribution of this work is showing that certain regularized equilibria do not possess the aforementioned non-correspondence problem---thus, computing them can be treated as perfect information problems. This result yields a simplified framework for decision-time planning in two-player zero-sum games, void of the unappealing properties that plague existing decision-time planning algorithms.

****

Is Attention All That NeRF Needs?

Mukund Varma T,Peihao Wang,Xuxi Chen,Tianlong Chen,Subhashini Venugopalan,Zhangyang Wang

We present Generalizable NeRF Transformer (GNT), a transformer-based architecture that reconstructs Neural Radiance Fields (NeRFs) and learns to render novel views on the fly from source views. While prior works on NeRFs optimize a scene representation by inverting a handcrafted rendering equation, GNT achieves neural representation and rendering that generalizes across scenes using transformers at two stages. (1) The view transformer leverages multi-view geometry as an inductive bias for attention-based scene representation, and predicts coordinate-aligned features by aggregating information from epipolar lines on the neighboring views. (2) The ray transformer renders novel views using attention to decode the features from the view transformer along the sampled points during ray marching. Our experiments demonstrate that when optimized on a single scene, GNT can successfully reconstruct NeRF without an explicit rendering formula due to the learned ray renderer. When trained on multiple scenes, GNT consistently achieves state-of-the-art performance when transferring to unseen scenes and outperform all other methods by ~10% on average. Our analysis of the learned attention maps to infer depth and occlusion indicate that attention enables learning a physically-grounded rendering. Our results show the promise of transformers as a universal modeling tool for graphics. Please refer to our project page for video results: https://vita-group.github.io/GNT/

****

Stochastic No-regret Learning for General Games with Variance Reduction

Yichi Zhou,Fang Kong,Shuai Li

We show that a stochastic version of optimistic mirror descent (OMD), a variant of mirror descent with recency bias, converges fast in general games. More specifically, with our algorithm, the individual regret of each player vanishes at a speed of $O(1/T^{3/4})$ and the sum of all players' regret vanishes at a speed of $O(1/T)$, which is an improvement upon the $O(1/\sqrt{T})$ convergence rate of prior stochastic algorithms, where T is the number of interaction rounds. Due to the advantage of stochastic methods in the computational cost, we significantly improve the time complexity over the deterministic algorithms to approximate coarse correlated equilibrium. To achieve lower time complexity, we equip the stochastic version of OMD in \cite{alacaoglu2021stochastic} with a novel low-variance Monte-Carlo estimator. Our algorithm extends previous works \cite{alacaoglu2021stochastic,carmon2019variance} from two-player zero-sum games to general games.

****

The Dark Side of AutoML: Towards Architectural Backdoor Search

Ren Pang,Changjiang Li,Zhaohan Xi,Shouling Ji,Ting Wang

This paper asks the intriguing question: is it possible to exploit neural architecture search (NAS) as a new attack vector to launch previously improbable attacks? Specifically, we present EVAS, a new attack that leverages NAS to find neural architectures with inherent backdoors and exploits such vulnerability using input-aware triggers. Compared with existing attacks, EVAS demonstrates many interesting properties: (i) it does not require polluting training data or perturbing model parameters; (ii) it is agnostic to downstream fine-tuning or even re-training from scratch; (iii) it naturally evades defenses that rely on inspecting model parameters or training data. With extensive evaluation on benchmark datasets

, we show that EVAS features high evasiveness, transferability, and robustness, thereby expanding the adversary's design spectrum. We further characterize the mechanisms underlying EVAS, which are possibly explainable by architecture-level ``shortcuts'' that recognize trigger patterns. This work showcases that NAS can be exploited in a harmful way to find architectures with inherent backdoor vulnerability. The code is available at https://github.com/ain-soph/nas_backdoor.

****

Generalization and Estimation Error Bounds for Model-based Neural Networks

Avner Shultzman,Eyar Azar,Miguel R. D. Rodrigues,Yonina C. Eldar

Model-based neural networks provide unparalleled performance for various tasks, such as sparse coding and compressed sensing problems. Due to the strong connection with the sensing model, these networks are interpretable and inherit prior structure of the problem. In practice, model-based neural networks exhibit higher generalization capability compared to ReLU neural networks. However, this phenomenon was not addressed theoretically. Here, we leverage complexity measures including the global and local Rademacher complexities, in order to provide upper bounds on the generalization and estimation errors of model-based networks. We show that the generalization abilities of model-based networks for sparse recovery outperform those of regular ReLU networks, and derive practical design rules that allow to construct model-based networks with guaranteed high generalization. We demonstrate through a series of experiments that our theoretical insights shed light on a few behaviours experienced in practice, including the fact that ISTA and ADMM networks exhibit higher generalization abilities (especially for small number of training samples), compared to ReLU networks.

****

Isometric Representations in Neural Networks Improve Robustness

Kosio Beshkov,Jonas Verhellen,Mikkel Elle Lepperød

Artificial and biological agents are unable to learn given completely random and unstructured data. The structure of data is encoded in the distance or similarity relationships between data points. In the context of neural networks, the neuronal activity within a layer forms a representation reflecting the transformation that the layer implements on its inputs. In order to utilize the structure in the data in a truthful manner, such representations should reflect the input distances and thus be continuous and isometric. Supporting this statement, recent findings in neuroscience propose that generalization and robustness are tied to neural representations being continuously differentiable. However, in machine learning, most algorithms lack robustness and are generally thought to rely on aspects of the data that differ from those that humans use, as is commonly seen in adversarial attacks.

During cross-entropy classification, the metric and structural properties of network representations are usually broken both between and within classes. This side effect from training can lead to instabilities under perturbations near locations where such structure is not preserved. One of the standard solutions to obtain robustness is to train specifically by introducing perturbations in the training data. This leads to networks that are particularly robust to specific training perturbations but not necessarily to general perturbations. While adding ad hoc regularization terms to improve robustness has become common practice, to our knowledge, forcing representations to preserve the metric structure of the input data as a stabilising mechanism has not yet been introduced.

In this work, we train neural networks to perform classification while simultaneously maintaining the metric structure within each class, leading to continuous and isometric within-class representations. We show that such network representations turn out to be a beneficial component for making accurate and robust inferences about the world. By stacking layers with this property we provide the community with an network architecture that facilitates hierarchical manipulation of internal neural representations. Finally, we verify that our isometric regularization term improves the robustness to adversarial attacks on MNIST.

****

TAN without a burn: Scaling laws of DP-SGD
Tom Sander,Pierre Stock,Alexandre Sablayrolles

Differentially Private methods for training Deep Neural Networks (DNNs) have progressed recently, in particular with the use of massive batches and aggregated data augmentations for a large number of steps. These techniques require much more compute than their non-private counterparts, shifting the traditional privacy-accuracy trade-off to a privacy-accuracy-compute trade-off and making hyper-parameter search virtually impossible for realistic scenarios. In this work, we decouple privacy analysis and experimental behavior of noisy training to explore the trade-off with minimal computational requirements. We first use the tools of Rényi Differential Privacy (RDP) to show that the privacy budget, when not overcharged, only depends on the total amount of noise (TAN) injected throughout training. We then derive scaling laws for training models with DP-SGD to optimize hyper-parameters with more than a $100\times$ reduction in computational budget. We apply the proposed method on CIFAR-10 and ImageNet and, in particular, strongly improve the state-of-the-art on ImageNet with a $+9$ points gain in accuracy for a privacy budget $\varepsilon=8$.

****

A sampling framework for value-based reinforcement learning
Frank Shih,Faming Liang

Value-based algorithms have achieved great successes in solving Reinforcement Learning problems via minimizing the mean squared Bellman error (MSBE). Temporal-difference (TD) algorithms such as Q-learning and SARSA often use stochastic gradient descent based optimization approaches to estimate the value function parameters, but fail to quantify their uncertainties. In our work, under the Kalman filtering paradigm, we establish a novel and scalable sampling framework based on stochastic gradient Markov chain Monte Carlo, which allows us to efficiently generate samples from the posterior distribution of deep neural network parameters. For TD-learning with both linear and nonlinear function approximation, we prove that the proposed algorithm converges to a stationary distribution, which allows us to measure uncertainties of the value function and its parameters.

****

The Curse of Low Task Diversity: On the Failure of Transfer Learning to Outperform MAML and their Empirical Equivalence
Brando Miranda,Patrick Yu,Yu-Xiong Wang,Oluwasanmi O Koyejo

Recently, it has been observed that a transfer learning solution might be all we need to solve many few-shot learning benchmarks -- thus raising important questions about when and how meta-learning algorithms should be deployed.
In this paper, we seek to clarify these questions by
1. proposing a novel metric -- the {\it diversity coefficient} -- to measure the diversity of tasks in a few-shot learning benchmark and
2. by comparing MAML and transfer learning under fair conditions (same architecture, same optimizer and all models trained to convergence).
Using the diversity coefficient, we show that the popular MiniImagenet and Cifar-fs few-shot learning benchmarks have low diversity.
This novel insight contextualizes claims that transfer learning solutions are better than meta-learned solutions in the regime of low diversity under a fair comparison.
Specifically, we empirically find that a low diversity coefficient correlates with a high similarity between transfer learning and Model-Agnostic Meta-Learning (MAML) learned solutions in terms of accuracy at meta-test time and classification layer similarity (using feature based distance metrics like SVCCA, PWCCA, CKA, and OPD).
To further support our claim, we find this meta-test accuracy holds even as the model size changes.
Therefore, we conclude that in the low diversity regime, MAML and transfer learning have equivalent meta-test performance when both are compared fairly.
We also hope our work inspires more thoughtful constructions and quantitative evaluations of meta-learning benchmarks in the future.

****

Bi-Stride Multi-Scale Graph Neural Network for Mesh-Based Physical Simulation

Yadi Cao,Menglei Chai,Minchen Li,Chenfanfu Jiang

Learning physical systems on unstructured meshes by flat Graph neural networks (GNNs) faces the challenge of modeling the long-range interactions due to the scaling complexity w.r.t. the number of nodes, limiting the generalization under mesh refinement. On regular grids, the convolutional neural networks (CNNs) with a U-net structure can resolve this challenge by efficient stride, pooling, and upsampling operations. Nonetheless, these tools are much less developed for graph neural networks (GNNs), especially when GNNs are employed for learning large-scale mesh-based physics. The challenges arise from the highly irregular meshes and the lack of effective ways to construct the multi-level structure without losing connectivity. Inspired by the bipartite graph determination algorithm, we introduce Bi-Stride Multi-Scale Graph Neural Network (BSMS-GNN) by proposing \textit{bi-stride} as a simple pooling strategy for building the multi-level GNN. \textit{Bi-stride} pools nodes by striding every other BFS frontier; it 1) works robustly on any challenging mesh in the wild, 2) avoids using a mesh generator at coarser levels, 3) avoids the spatial proximity for building coarser levels, and 4) uses non-parametrized aggregating/returning instead of MLPs during pooling and unpooling. Experiments show that our framework significantly outperforms the state-of-the-art method's computational efficiency in representative physics-based simulation cases.

****

Spatially Resolved Temporal Networks: Online Unsupervised Representation Learning of High Frequency Time Series

Faris Faried Gulamali,Ashwin Shreekant Sawant,Ira Hofer,Matt Levin,Karandeep Singh,Benjamin S Glicksberg,Girish N Nadkarni

Univariate high-frequency time series are dominant data sources for many medical, economic and environmental applications. In many of these domains, the time series are tied to real-time changes in state. In the intensive care unit, for example, changes in an electrocardiogram signal can indicate a heart attack, and intracranial pressure waveforms can indicate whether a patient is developing decreased blood perfusion to the brain. However, most representation learning to resolve states is conducted in an offline, batch-dependent manner. In high frequency time-series, high intra-state and inter-sample variability makes offline, batch-dependent learning a relatively difficult task. Hence, we propose Spatial Resolved Temporal Networks (SpaRTeN), a novel composite deep learning model for online, unsupervised representation learning through a spatially constrained latent space. We simultaneously train two distinct blocks: a recurrent neural network ensemble f_R that captures states in high frequency time series, and a spatial block f_S that spatially resolves state changes from the predictions generated by f_R. The spatial block f_S identifies the block in f_R that best fits the current state of the time series, and the training procedure for f_R optimizes that block. This procedure corresponds to a minimax framework. When f_S and f_R are deep neural networks, the entire system can be trained via back-propagation. Finally, we demonstrate the application of this framework to online forecasting and interpretable, zero-shot clustering. We compare and demonstrate that SpaRTeN outperforms spectral clustering and a Gaussian mixture model.

****

ChordMixer: A Scalable Neural Attention Model for Sequences with Different Length

Ruslan Khalitov,Tong Yu,Lei Cheng,Zhirong Yang

Sequential data naturally have different lengths in many domains, with some very long sequences. As an important modeling tool, neural attention should capture long-range interaction in such sequences. However, most existing neural attention models admit only short sequences, or they have to employ chunking or padding to enforce a constant input length. Here we propose a simple neural network building block called ChordMixer which can model the attention for long sequences with variable lengths. Each ChordMixer block consists of a position-wise rotation layer without learnable parameters and an element-wise MLP layer. Repeatedly applying such blocks forms an effective network backbone that mixes the input signa ls towards the learning targets. We have tested ChordMixer on the synthetic adding problem, long document classification, and DNA sequence-based taxonomy classification. The experiment results show that our method substantially outperforms other neural attention models.

****

Boosting Adversarial Transferability using Dynamic Cues

Muzammal Naseer,Ahmad Mahmood,Salman Khan,Fahad Khan

The transferability of adversarial perturbations between image models has been extensively studied. In this case, an attack is generated from a known surrogate \eg, the ImageNet trained model, and transferred to change the decision of an unknown (black-box) model trained on an image dataset. However, attacks generated from image models do not capture the dynamic nature of a moving object or a changing scene due to a lack of temporal cues within image models. This leads to reduced transferability of adversarial attacks from representation-enriched \emph{image} models such as Supervised Vision Transformers (ViTs), Self-supervised ViTs (\eg, DINO), and Vision-language models (\eg, CLIP) to black-box \emph{video} models. In this work, we induce dynamic cues within the image models without sacrificing their original performance on images. To this end, we optimize \emph{temporal prompts} through frozen image models to capture motion dynamics. Our temporal prompts are the result of a learnable transformation that allows optimizing for temporal gradients during an adversarial attack to fool the motion dynamics. Specifically, we introduce spatial (image) and temporal (video) cues within the same source model through task-specific prompts. Attacking such prompts maximizes the adversarial transferability from image-to-video and image-to-image models using the attacks designed for image models. As an example, an iterative attack launched from image model Deit-B with temporal prompts reduces generalization (top1 \% accuracy) of a video model by 35\% on Kinetics-400. Our approach also improves adversarial transferability to image models by 9\% on ImageNet w.r.t the current state-of-the-art approach. Our attack results indicate that the attacker does not need specialized architectures, \eg, divided space-time attention, 3D convolutions, or multi-view convolution networks for different data modalities. Image models are effective surrogates to optimize an adversarial attack to fool black-box models in a changing environment over time. Code is available at \url{https://bit.ly/3Xd9gRQ}

****

Static Prediction of Runtime Errors by Learning to Execute Programs with External Resource Descriptions

David Bieber,Rishab Goel,Dan Zheng,Hugo Larochelle,Daniel Tarlow

The execution behavior of a program often depends on external resources, such as program inputs or file contents, and so the program cannot be run in isolation. Nevertheless, software developers benefit from fast iteration loops where automated tools identify errors as early as possible, even before programs can be compiled and run. This presents an interesting machine learning challenge: can we predict runtime errors in a "static" setting, where program execution is not possible? Here, we introduce a competitive programming dataset and task for predicting runtime errors, which we show is difficult for generic models like Transformers. We approach this task by developing an interpreter-inspired architecture with an inductive bias towards mimicking program executions, which models exception handling and "learns to execute" descriptions of external resources. Surprisingly, we show that the model can also predict the locations of errors, despite being trained only on labels indicating error presence or absence and kind. In total, we present a practical and difficult-yet-approachable challenge problem related to learning program execution behavior and we demonstrate promising new capabilities of interpreter-inspired machine learning models for code.

****

Matching receptor to odorant with protein language and graph neural networks

Matej Hladiš,Maxence Lalis,Sebastien Fiorucci,Jérémie Topin

Odor perception in mammals is triggered by interactions between volatile organic compounds and a subset of hundreds of proteins called olfactory receptors (ORs). Molecules activate these receptors in a complex combinatorial coding allowing mammals to discriminate a vast number of chemical stimuli. Recently, ORs have gained attention as new therapeutic targets following the discovery of their involvement in other physiological processes and diseases. To date, predicting molecule-induced activation for ORs is highly challenging since 43% of ORs have no identified active compound. In this work, we combine [CLS] token from protBERT with a molecular graph and propose a tailored GNN architecture incorporating inductive biases from the protein-molecule binding. We abstract the biological process of protein-molecule activation as the injection of a molecule into a protein-specific environment. On a newly gathered dataset of 46 700 OR-molecule pairs, this model outperforms state-of-the-art models on drug-target interaction prediction as well as standard GNN baselines. Moreover, by incorporating non-bonded interactions the model is able to work with mixtures of compounds. Finally, our predictions reveal a similar activation pattern for molecules within a given odor family, which is in agreement with the theory of combinatorial coding in olfaction.

****

How does overparametrization affect performance on minority groups?

Subha Maity,Saptarshi Roy,Songkai Xue,Mikhail Yurochkin,Yuekai Sun

The benefits of overparameterization for the overall performance of modern machine learning (ML) models are well known. However, the effect of overparameterization at a more granular level of data subgroups is less understood. Recent empirical studies demonstrate encouraging results: (i) when groups are not known, overparameterized models trained with empirical risk minimization (ERM) perform better on minority groups; (ii) when groups are known, ERM on data subsampled to equalize group sizes yields state-of-the-art worst-group-accuracy in the overparameterized regime. In this paper, we complement these empirical studies with a theoretical investigation of the risk of overparameterized random feature models on minority groups. In a setting in which the regression functions for the majority and minority groups are different, we show that overparameterization always improves minority group performance.

****

Federated Training of Dual Encoding Models on Small Non-IID Client Datasets

Raviteja Vemulapalli,Warren Richard Morningstar,Philip Andrew Mansfield,Hubert Eichner,Karan Singhal,Arash Afkanpour,Bradley Green

Dual encoding models that encode a pair of inputs are widely used for representation learning. Many approaches train dual encoding models by maximizing agreement between pairs of encodings on centralized training data. However, in many scenarios, datasets are inherently decentralized across many clients (user devices or organizations) due to privacy concerns, motivating federated learning. In this work, we focus on federated training of dual encoding models on decentralized data composed of many small, non-IID (independent and identically distributed) client datasets. We show that existing approaches that work well in centralized settings perform poorly when naively adapted to this setting using federated averaging. We observe that, we can simulate large-batch loss computation on individual clients for loss functions that are based on encoding statistics. Based on this insight, we propose a novel federated training approach, Distributed Cross Correlation Optimization (DCCO), which trains dual encoding models using encoding statistics aggregated across clients, without sharing individual data samples. Our experimental results on two datasets demonstrate that the proposed DCCO approach outperforms federated variants of existing approaches by a large margin.

****

Offline Policy Comparison with Confidence: Benchmarks and Baselines

Anurag Koul,Mariano Phielipp,Alan Fern

Decision makers often wish to use offline historical data to compare sequential-action policies at various world states. Importantly, computational tools should produce confidence values for such offline policy comparison (OPC) to account for statistical variance and limited data coverage. Nevertheless, there is little work that directly evaluates the quality of confidence values for OPC. In this work, we address this issue by creating benchmarks for OPC with Confidence (OPCC), derived by adding sets of policy comparison queries to datasets from offline reinforcement learning. In addition, we present an empirical evaluation of the "risk versus coverage" trade-off for a class of model-based baselines. In particular, the baselines learn ensembles of dynamics models, which are used in various ways to produce simulations for answering queries with confidence values. While our results suggest advantages for certain baseline variations, there appears to be significant room for improvement in future work.

****

SGDA with shuffling: faster convergence for nonconvex-P■ minimax optimization
Hanseul Cho,Chulhee Yun

Stochastic gradient descent-ascent (SGDA) is one of the main workhorses for solving finite-sum minimax optimization problems. Most practical implementations of SGDA randomly reshuffle components and sequentially use them (i.e., without-replacement sampling); however, there are few theoretical results on this approach for minimax algorithms, especially outside the easier-to-analyze (strongly-)monotone setups. To narrow this gap, we study the convergence bounds of SGDA with random reshuffling (SGDA-RR) for smooth nonconvex-nonconcave objectives with Polyak-{\L}ojasiewicz (P{\L}) geometry. We analyze both simultaneous and alternating SGDA-RR for nonconvex-P{\L} and primal-P{\L}-P{\L} objectives, and obtain convergence rates faster than with-replacement SGDA. Our rates extend to mini-batch SGDA-RR, recovering known rates for full-batch gradient descent-ascent (GDA). Lastly, we present a comprehensive lower bound for GDA with an arbitrary step-size ratio, which matches the full-batch upper bound for the primal-P{\L}-P{\L} case.

****

NTFields: Neural Time Fields for Physics-Informed Robot Motion Planning
Ruiqi Ni,Ahmed H Qureshi

Neural Motion Planners (NMPs) have emerged as a promising tool for solving robot navigation tasks in complex environments. However, these methods often require expert data for learning, which limits their application to scenarios where data generation is time-consuming. Recent developments have also led to physics-informed deep neural models capable of representing complex dynamical Partial Differential Equations (PDEs). Inspired by these developments, we propose Neural Time Fields (NTFields) for robot motion planning in cluttered scenarios. Our framework represents a wave propagation model generating continuous arrival time to find path solutions informed by a nonlinear first-order PDE called Eikonal Equation. We evaluate our method in various cluttered 3D environments, including the Gibson dataset, and demonstrate its ability to solve motion planning problems for 4-DOF and 6-DOF robot manipulators where the traditional grid-based Eikonal planners often face the curse of dimensionality. Furthermore, the results show that our method exhibits high success rates and significantly lower computational times than the state-of-the-art methods, including NMPs that require training data from classical planners.

****

MOAT: Alternating Mobile Convolution and Attention Brings Strong Vision Models
Chenglin Yang,Siyuan Qiao,Qihang Yu,Xiaoding Yuan,Yukun Zhu,Alan Yuille,Hartwig Adam,Liang-Chieh Chen

This paper presents MOAT, a family of neural networks that build on top of MObile convolution (i.e., inverted residual blocks) and ATtention. Unlike the current works that stack separate mobile convolution and transformer blocks, we effectively merge them into a MOAT block. Starting with a standard Transformer block, we replace its multi-layer perceptron with a mobile convolution block, and further reorder it before the self-attention operation. The mobile convolution block not only enhances the network representation capacity, but also produces better downsampled features. Our conceptually simple MOAT networks are surprisingly effective, achieving 89.1% / 81.5% top-1 accuracy on ImageNet-1K / ImageNet-1K-V2 with ImageNet-22K pretraining. Additionally, MOAT can be seamlessly applied to downstream tasks that require large resolution inputs by simply converting the global attention to window attention. Thanks to the mobile convolution that effectively exchanges local information between pixels (and thus cross-windows), MOAT does not need the extra window-shifting mechanism. As a result, on COCO object detection, MOAT achieves 59.2% AP$^{\text{box}}$ with 227M model parameters (single

-scale inference, and hard NMS), and on ADE20K semantic segmentation, MOAT attains 57.6% mIoU with 496M model parameters (single-scale inference). Finally, the tiny-MOAT family, obtained by simply reducing the channel sizes, also surprisingly outperforms several mobile-specific transformer-based models on ImageNet. The tiny-MOAT family is also benchmarked on downstream tasks, serving as a baseline for the community. We hope our simple yet effective MOAT will inspire more seamless integration of convolution and self-attention. Code is publicly available.
****

Part-Based Models Improve Adversarial Robustness
Chawin Sitawarin,Kornrapat Pongmala,Yizheng Chen,Nicholas Carlini,David Wagner
We show that combining human prior knowledge with end-to-end learning can improve the robustness of deep neural networks by introducing a part-based model for object classification. We believe that the richer form of annotation helps guide neural networks to learn more robust features without requiring more samples or larger models. Our model combines a part segmentation model with a tiny classifier and is trained end-to-end to simultaneously segment objects into parts and then classify the segmented object. Empirically, our part-based models achieve both higher accuracy and higher adversarial robustness than a ResNet-50 baseline on all three datasets. For instance, the clean accuracy of our part models is up to 15 percentage points higher than the baseline's, given the same level of robustness. Our experiments indicate that these models also reduce texture bias and yield better robustness against common corruptions and spurious correlations. The code is publicly available at https://github.com/chawins/adv-part-model.
****

PGrad: Learning Principal Gradients For Domain Generalization
Zhe Wang,Jake Grigsby,Yanjun Qi
Machine learning models fail to perform when facing out-of-distribution (OOD) domains, a challenging task known as domain generalization (DG). In this work, we develop a novel DG training strategy, we call PGrad, to learn a robust gradient direction, improving models' generalization ability on unseen domains. The proposed gradient aggregates the principal directions of a sampled roll-out optimization trajectory that measures the training dynamics across all training domains. PGrad gradient design forces the DG training to ignore domain-dependent noise signals and updates all training domains with a robust direction covering main components of parameter dynamics. We further improve PGrad via bijection-based computational refinement and directional plus length-based calibrations. Our theoretical proof connects PGrad to the spectral analysis of Hessian in training neural networks. Experiments on DomainBed and WILDS benchmarks demonstrate that our approach effectively enables robust DG optimization and leads to smoothly decreased loss curves. Empirically, PGrad achieves competitive results across seven datasets, demonstrating its efficacy across both synthetic and real-world distributional shifts.
****

Learning Efficient Models From Few Labels By Distillation From Multiple Tasks
Kenneth Borup,Cheng Perng Phoo,Bharath Hariharan
We address the challenge of getting efficient yet accurate recognition systems that can be trained with limited labels. Many specialized applications of computer vision (e.g. analyzing X-rays or satellite images) have severe resource constraints both during training and inference. While transfer learning is an effective solution for training on small labeled datasets it still often requires a large base model for fine-tuning. In this paper we present a weighted multi-source distillation method; we distill multiple (diverse) source models trained on different domains, weighted by their relevance for the target task, into a single efficient model using limited labeled data. When the goal is accurate recognition under computational constraints, our approach outperforms both transfer learning from strong ImageNet initializations as well as state-of-the-art semi-supervised techniques such as FixMatch. When averaged over 8 diverse target tasks our method outperform the baselines by 5.6%-points and 4.5%-points, respectively.
****

Extremely Simple Activation Shaping for Out-of-Distribution Detection

Andrija Djurisic,Nebojsa Bozanic,Arjun Ashok,Rosanne Liu

The separation between training and deployment of machine learning models implies that not all scenarios encountered in deployment can be anticipated during training, and therefore relying solely on advancements in training has its limits. Out-of-distribution (OOD) detection is an important area that stress-tests a model’s ability to handle unseen situations: Do models know when they don’t know? Existing OOD detection methods either incur extra training steps, additional data or make nontrivial modifications to the trained network. In contrast, in this work, we propose an extremely simple, post-hoc, on-the-fly activation shaping method, ASH, where a large portion (e.g. 90%) of a sample’s activation at a late layer is removed, and the rest (e.g. 10%) simplified or lightly adjusted. The shaping is applied at inference time, and does not require any statistics calculated from training data. Experiments show that such a simple treatment enhances in-distribution and out- of-distribution sample distinction so as to allow state-of-the-art OOD detection on ImageNet, and does not noticeably deteriorate the in-distribution accuracy. Video, animation and code can be found at: https://andrijazz.github.io/ash.

****

ZiCo: Zero-shot NAS via inverse Coefficient of Variation on Gradients

Guihong Li,Yuedong Yang,Kartikeya Bhardwaj,Radu Marculescu

Neural Architecture Search (NAS) is widely used to automatically obtain the neural network with the best performance among a large number of candidate architectures. To reduce the search time, zero-shot NAS aims at designing training-free proxies that can predict the test performance of a given architecture. However, as shown recently, none of the zero-shot proxies proposed to date can actually work consistently better than a naive proxy, namely, the number of network parameters (#Params). To improve this state of affairs, as the main theoretical contribution, we first reveal how some specific gradient properties across different samples impact the convergence rate and generalization capacity of neural networks. Based on this theoretical analysis, we propose a new zero-shot proxy, ZiCo, the first proxy that works consistently better than #Params. We demonstrate that ZiCo works better than State-Of-The-Art (SOTA) proxies on several popular NAS-Benchmarks (NASBench101, NATSBench-SSS/TSS, TransNASBench-101) for multiple applications (e.g., image classification/reconstruction and pixel-level prediction). Finally, we demonstrate that the optimal architectures found via ZiCo are as competitive as the ones found by one-shot and multi-shot NAS methods, but with much less search time. For example, ZiCo-based NAS can find optimal architectures with 78.1%, 79.4%, and 80.4% test accuracy under inference budgets of 450M, 600M, and 1000M FLOPs, respectively, on ImageNet within 0.4 GPU days. Our code is available at https://github.com/SLDGroup/ZiCo.

****

Statistical Guarantees for Consensus Clustering

Zhixin Zhou,Gautam Dudeja,Arash A Amini

Consider the problem of clustering n objects. One can apply multiple algorithms to produce N potentially different clustersings of the same objects, that is, partitions of the n objects into K groups. Even a single randomized algorithm can output different clusterings. This often happens when one samples from the posterior of a Bayesian model, or runs multiple MCMC chains from random initializations. A natural task is then to form a consensus among these different clusterings. The challenge in an unsupervised setting is that the optimal matching between clusters of different inputs is unknown. We model this problem as finding a barycenter (also known as Fr\'{e}chet mean) relative to the misclassification rate. We show that by lifting the problem to the space of association matrices, one can derive aggregation algorithms that circumvent the knowledge of the optimal matchings. We analyze the statistical performance of aggregation algorithms under a stochastic label perturbation model, and show that a K-means type algorithm followed by a local refinement step can achieve near optimal performance, with a rate that decays exponentially fast in N. Numerical experiments show t he effectiveness of the proposed methods.

****

Perceive, Ground, Reason, and Act: A Benchmark for General-purpose Visual Representation

Jiangyong Huang,William Yicheng Zhu,Baoxiong Jia,Zan Wang,Xiaojian Ma,Qing Li,Siyuan Huang

Current computer vision models, unlike the human visual system, cannot yet achieve general-purpose visual understanding. Existing efforts at general vision models are limited to a narrow range of tasks and offer no overarching framework to perform visual tasks holistically. We present a new comprehensive benchmark, General-purpose Visual Understanding Evaluation (G-VUE), covering the full spectrum of visual cognitive abilities with four disjoint functional domains — Perceive, Ground, Reason, and Act. The four domains are embodied in 11 carefully curated tasks, from 3D reconstruction to visual reasoning and manipulation. Along with the benchmark, we provide a general encoder-decoder framework for the tasks in G-VUE, to accommodate arbitrary visual representations on all 11 tasks. With our benchmark and framework, we evaluate 7 typical visual representations and observe that (1) transformer and more data empirically lead to more general-purpose, (2) language plays a significant role in learning versatile visual representation, and (3) correlations indicate a subtle constituent among tasks despite the distinctions, which could be evidence of general-purpose. We argue that instead of pursuing general-purpose vision models by end-to-end multi-task training, it is more reasonable to evaluate and investigate representations, which helps digest emerging pre-trained vision models and hopefully shed light on general intelligence.

****

Expressive Monotonic Neural Networks

Niklas Nolte,Ouail Kitouni,Mike Williams

The monotonic dependence of the outputs of a neural network on some of its inputs is a crucial inductive bias in many scenarios where domain knowledge dictates such behavior. This is especially important for interpretability and fairness considerations. In a broader context, scenarios in which monotonicity is important can be found in finance, medicine, physics, and other disciplines. It is thus desirable to build neural network architectures that implement this inductive bias provably. In this work, we propose a weight-constrained architecture with a single residual connection to achieve exact monotonic dependence in any subset of the inputs. The weight constraint scheme directly controls the Lipschitz constant of the neural network and thus provides the additional benefit of robustness. Compared to currently existing techniques used for monotonicity, our method is simpler in implementation and in theory foundations, has negligible computational overhead, is guaranteed to produce monotonic dependence, and is highly expressive. We show how the algorithm is used to train powerful, robust, and interpretable discriminators that achieve competitive performance compared to current state-of-the-art methods across various benchmarks, from social applications to the classification of the decays of subatomic particles produced at the CERN Large Hadron Collider.

****

Active Image Indexing

Pierre Fernandez,Matthijs Douze,Herve Jegou,Teddy Furon

Image copy detection and retrieval from large databases leverage two components. First, a neural network maps an image to a vector representation, that is relatively robust to various transformations of the image. Second, an efficient but approximate similarity search algorithm trades scalability (size and speed) against quality of the search, thereby introducing a source of error.

This paper improves the robustness of image copy detection with active indexing, that optimizes the interplay of these two components. We reduce the quantization loss of a given image representation by making imperceptible changes to the image before its release. The loss is back-propagated through the deep neural network back to the image, under perceptual constraints. These modifications make the image more retrievable.

Our experiments show that the retrieval and copy detection of activated images is significantly improved. For instance, activation improves by $+40\%$ the Recall1@1 on various image transformations, and for several popular indexing structures based on product quantization and locality sensitivity hashing.

****

Towards Explaining Distribution Shifts

Sean Kulinski,David I. Inouye

Distribution shift can have fundamental consequences such as signaling a change in the operating environment or significantly reducing the accuracy of downstream models. Thus, understanding such distribution shifts is critical for examining and hopefully mitigating the effect of such a shift. Most prior work has focused on merely detecting if a shift has occurred and assumes any detected shift can be understood and handled appropriately by a human operator. We hope to aid in these manual mitigation tasks by explaining the distribution shift using interpretable transportation maps from the original distribution to the shifted one. We derive our interpretable mappings from a relaxation of the optimal transport problem, where the candidate mappings are restricted to belong to a set of interpretable mappings. We then use quintessential examples of distribution shift in simulated and real-world cases to showcase how our explanatory mappings provide a better balance between detail and interpretability than the de facto standard mean shift explanation by both visual inspection and our PercentExplained metric.

****

Perturbation Analysis of Neural Collapse

Tom Tirer,Haoxiang Huang,Jonathan Niles-Weed

Training deep neural networks for classification often includes minimizing the training loss beyond the zero training error point. In this phase of training, a “neural collapse” behavior has been observed: the variability of features (outputs of the penultimate layer) of within-class samples decreases and the mean features of different classes approach a certain tight frame structure. Recent works analyze this behavior via idealized unconstrained features models where all the minimizers exhibit exact collapse. However, with practical networks and datasets, the features typically do not reach exact collapse, e.g., because deep layers cannot arbitrarily modify intermediate features that are far from being collapsed. In this paper, we propose a richer model that can capture this phenomenon by forcing the features to stay in the vicinity of a predefined features matrix (e.g., intermediate features). We explore the model in the small vicinity case via perturbation analysis and establish results that cannot be obtained by the previously studied models. For example, we prove reduction in the within-class variability of the optimized features compared to the predefined input features (via analyzing gradient flow on the “central-path” with minimal assumptions), analyze the minimizers in the near-collapse regime, and provide insights on the effect of regularization hyperparameters on the closeness to collapse. We support our theory with experiments in practical deep learning settings.

****

Learning Simultaneous Navigation and Construction in Grid Worlds

Wenyu Han,Haoran Wu,Eisuke Hirota,Alexander Gao,Lerrel Pinto,Ludovic Righetti,Chen Feng

We propose to study a new learning task, mobile construction, to enable an agent to build designed structures in 1/2/3D grid worlds while navigating in the same evolving environments. Unlike existing robot learning tasks such as visual navigation and object manipulation, this task is challenging because of the interdependence between accurate localization and strategic construction planning. In pursuit of generic and adaptive solutions to this partially observable Markov decision process (POMDP) based on deep reinforcement learning (RL), we design
a Deep Recurrent Q-Network (DRQN) with explicit recurrent position estimation in this dynamic grid world. Our extensive experiments show that pre-training this position estimation module before Q-learning can significantly improve the construction performance measured by the intersection-over-union score, achieving the best results in our benchmark of various baselines including model-free and model-based RL, a handcrafted SLAM-based policy, and human players. Our code is ava

ilable at: https://ai4ce.github.io/SNAC/.

****

Learning to CROSS exchange to solve min-max vehicle routing problems

Minjun Kim,Junyoung Park,Jinkyoo Park

CROSS exchange (CE), a meta-heuristic that solves various vehicle routing problems (VRPs), improves the solutions of VRPs by swapping the sub-tours of the vehicles. Inspired by CE, we propose Neuro CE (NCE), a fundamental operator of \textit{learned} meta-heuristic, to solve various min-max VRPs while overcoming the limitations of CE, i.e., the expensive $\mathcal{O}(n^4)$ search cost. NCE employs graph neural network to predict the cost-decrements (i.e., results of CE searches) and utilizes the predicted cost-decrements to guide the selection of sub-tours for swapping, while reducing the search cost to $\mathcal{O}(n^2)$. As the learning objective of NCE is to predict the cost-decrement, the training can be simply done in a supervised fashion, whose training samples can be easily collected. Despite the simplicity of NCE, numerical results show that the NCE trained with min-max flexible multi-depot VRP (min-max FMDVRP) outperforms the meta-heuristic baselines. More importantly, it significantly outperforms the neural baselines when solving distinctive special cases of min-max FMDVRP (e.g., min-max MDVRP, min-max mTSP, min-max CVRP) without additional training.

****

PandA: Unsupervised Learning of Parts and Appearances in the Feature Maps of GANs

James Oldfield,Christos Tzelepis,Yannis Panagakis,Mihalis Nicolaou,Ioannis Patras

Recent advances in the understanding of Generative Adversarial Networks (GANs) have led to remarkable progress in visual editing and synthesis tasks, capitalizing on the rich semantics that are embedded in the latent spaces of pre-trained GANs. However, existing methods are often tailored to specific GAN architectures and are limited to either discovering global semantic directions that do not facilitate localized control, or require some form of supervision through manually provided regions or segmentation masks. In this light, we present an architecture-agnostic approach that jointly discovers factors representing spatial parts and their appearances in an entirely unsupervised fashion. These factors are obtained by applying a semi-nonnegative tensor factorization on the feature maps, which in turn enables context-aware local image editing with pixel-level control. In addition, we show that the discovered appearance factors correspond to saliency maps that localize concepts of interest, without using any labels. Experiments on a wide range of GAN architectures and datasets show that, in comparison to the state of the art, our method is far more efficient in terms of training time and, most importantly, provides much more accurate localized control. Our code is available at: https://github.com/james-oldfield/PandA.

****

Compositional Law Parsing with Latent Random Functions

Fan Shi,Bin Li,Xiangyang Xue

Human cognition has compositionality. We understand a scene by decomposing the scene into different concepts (e.g., shape and position of an object) and learning the respective laws of these concepts, which may be either natural (e.g., laws of motion) or man-made (e.g., laws of a game). The automatic parsing of these laws indicates the model's ability to understand the scene, which makes law parsing play a central role in many visual tasks. This paper proposes a deep latent variable model for Compositional LAw Parsing (CLAP), which achieves the human-like compositionality ability through an encoding-decoding architecture to represent concepts in the scene as latent variables. CLAP employs concept-specific latent random functions instantiated with Neural Processes to capture the law of concepts. Our experimental results demonstrate that CLAP outperforms the baseline methods in multiple visual tasks such as intuitive physics, abstract visual reasoning, and scene representation. The law manipulation experiments illustrate CLAP's interpretability by modifying specific latent random functions on samples. For example, CLAP learns the laws of position-changing and appearance constancy from the moving balls in a scene, making it possible to exchange laws between sampl es or compose existing laws into novel laws.

****

Pink Noise Is All You Need: Colored Noise Exploration in Deep Reinforcement Learning

Onno Eberhard,Jakob Hollenstein,Cristina Pinneri,Georg Martius

In off-policy deep reinforcement learning with continuous action spaces, exploration is often implemented by injecting action noise into the action selection process. Popular algorithms based on stochastic policies, such as SAC or MPO, inject white noise by sampling actions from uncorrelated Gaussian distributions. In many tasks, however, white noise does not provide sufficient exploration, and temporally correlated noise is used instead. A common choice is Ornstein-Uhlenbeck (OU) noise, which is closely related to Brownian motion (red noise). Both red noise and white noise belong to the broad family of colored noise. In this work, we perform a comprehensive experimental evaluation on MPO and SAC to explore the effectiveness of other colors of noise as action noise. We find that pink noise, which is halfway between white and red noise, significantly outperforms white noise, OU noise, and other alternatives on a wide range of environments. Thus, we recommend it as the default choice for action noise in continuous control.

****

LilNetX: Lightweight Networks with EXtreme Model Compression and Structured Sparsification

Sharath Girish,Kamal Gupta,Saurabh Singh,Abhinav Shrivastava

We introduce LilNetX, an end-to-end trainable technique for neural networks that enables learning models with specified accuracy-rate-computation trade-off. Prior works approach these problems one at a time and often require post-processing or multistage training which become less practical and do not scale very well for large datasets or architectures. Our method constructs a joint training objective that penalizes the self information of network parameters in a latent representation space to encourage small model size while also introducing priors to increase structured sparsity in the parameter space to reduce computation. When compared with existing state-of-the-art model compression methods, we achieve up to 50% smaller model size and 98% model sparsity on ResNet-20 on the CIFAR-10 dataset as well as 37% smaller model size and 71% structured sparsity on ResNet-50 trained on ImageNet while retaining the same accuracy as those methods. We show that the resulting sparsity can improve the inference time of the models by almost 1.8 times the dense ResNet-50 baseline model. Code is available at https://github.com/Sharath-girish/LilNetX.

****

First-order Context-based Adaptation for Generalizing to New Dynamical Systems

Junyoung Park,Federico Berto,Arec Jamgochian,Mykel Kochenderfer,Jinkyoo Park

In this paper, we propose FOCA (First-Order Context-based Adaptation), a learning framework to model sets of systems governed by common but unknown laws that differentiate themselves by their context. Inspired by classical modeling-and-identification approaches, FOCA learns to represent the common law through shared parameters and relies on online optimization to compute system-specific context. Due to the online optimization-based context inference, the training of FOCA involves a bi-level optimization problem. To train FOCA efficiently, we utilize an exponential moving average (EMA)-based method that allows for fast training using only first-order derivatives. We test FOCA on polynomial regression and time-series prediction tasks composed of three ODEs and one PDE, empirically finding it outperforms baselines.

****

CBP-QSNN: Spiking Neural Networks Quantized Using Constrained Backpropagation

Donghyung Yoo,Doo Seok Jeong

Spiking Neural Networks (SNNs) support sparse event-based data processing at high power efficiency when implemented in event-based neuromorphic processors. However, the limited on-chip memory capacity of neuromorphic processors strictly delimits the depth and width of SNNs implemented. A direct solution is the use of q uantized SNNs (QSNNs) in place of SNNs with FP32 weights. To this end, we propose a method to quantize the weights using constrained backpropagation (CBP) with the Lagrangian function (conventional loss function plus well-defined weight-constraint functions) as an objective function. This work utilizes CBP as a post-training algorithm for deep SNNs pre-trained using various state-of-the-art methods including direct training (TSSL-BP, STBP, and surrogate gradient) and DNN-to-SNN conversion (SNN-Calibration), validating CBP as a general framework for QSNNs. CBP-QSNNs highlight their high accuracy insomuch as the degradation of accuracy on CIFAR-10, DVS128 Gesture, and CIFAR10-DVS in the worst case is less than 1\%. Particularly, CBP-QSNNs for SNN-Calibration-pretrained SNNs on CIFAR-100 highlight an unexpected large increase in accuracy by 3.72\% while using small weight-memory (3.5\% of the FP32 case).

****

Leveraging the Third Dimension in Contrastive Learning

Sumukh K Aithal,Anirudh Goyal,Alex Lamb,Yoshua Bengio,Michael Curtis Mozer

Self-Supervised Learning (SSL) methods operate on unlabeled data to learn robust representations useful for downstream tasks. Most SSL methods rely on augmentations obtained by transforming the 2D image pixel map. These augmentations ignore the fact that biological vision takes place in an immersive three-dimensional, temporally contiguous environment, and that low-level biological vision relies heavily on depth cues. Using a signal provided by a pretrained state-of-the-art RGB-to-depth model (the Depth Prediction Transformer, Ranftl et al., 2021), we explore two distinct approaches to incorporating depth signals into the SSL framework. First, we evaluate contrastive learning using an RGB+depth input representation. Second, we use the depth signal to generate novel views from slightly different camera positions, thereby producing a 3D augmentation for contrastive learning. We evaluate these two approaches on three different SSL methods---BYOL, SimSiam, and SwAV---using ImageNette (10 class subset of ImageNet) and ImageNet-100. We find that both approaches to incorporating depth signals improve the robustness and generalization of the baseline SSL methods, though the first approach (with depth-channel concatenation) is superior.

****

STaSy: Score-based Tabular data Synthesis

Jayoung Kim,Chaejeong Lee,Noseong Park

Tabular data synthesis is a long-standing research topic in machine learning. Many different methods have been proposed over the past decades, ranging from statistical methods to deep generative methods. However, it has not always been successful due to the complicated nature of real-world tabular data. In this paper, we present a new model named $\textbf{S}$core-based $\textbf{Ta}$bular data $\textbf{Sy}$nthesis ($\texttt{STaSy}$) and its training strategy based on the paradigm of score-based generative modeling. Despite the fact that score-based generative models have resolved many issues in generative models, there still exists room for improvement in tabular data synthesis. Our proposed training strategy includes a self-paced learning technique and a fine-tuning strategy, which further increases the sampling quality and diversity by stabilizing the denoising score matching training. Furthermore, we also conduct rigorous experimental studies in terms of the generative task trilemma: sampling quality, diversity, and time. In our experiments with 15 benchmark tabular datasets and 7 baselines, our method outperforms existing methods in terms of task-dependant evaluations and diversity.

****

REDUCING OVERSMOOTHING IN GRAPH NEURAL NETWORKS BY CHANGING THE ACTIVATION FUNCTION

Dimitrios Kelesis,Dimitrios Vogiatzis,Georgios Katsimpras,Dimitris Fotakis,Georgios Paliouras

The performance of Graph Neural Networks (GNNs) deteriorates as the depth of the network increases. That performance drop is mainly attributed to oversmoothing, which leads to similar node representations through repeated graph convolutions. We show that in deep GNNs the activation function plays a crucial role in over smoothing. We explain theoretically why this is the case and propose a simple modification to the slope of ReLU to reduce oversmoothing. The proposed approach enables deep architectures without the need to change the network architecture or to add residual connections. We verify the theoretical results experimentally and further show that deep networks, which do not suffer from oversmoothing are beneficial in the presence of the “cold start” problem, i.e. when there is no feature information about unlabeled nodes.

****

Visual Prompt Tuning For Test-time Domain Adaptation

Yunhe Gao,Xingjian Shi,Yi Zhu,Hao Wang,Zhiqiang Tang,Xiong Zhou,Mu Li,Dimitris N. Metaxas

Models should have the ability to adapt to unseen data during test-time to avoid performance drops caused by inevitable distribution shifts in real-world deployment scenarios. In this work, we tackle the practical yet challenging test-time adaptation (TTA) problem, where a model adapts to the target domain without accessing the source data. We propose a simple recipe called data-efficient prompt tuning (DePT) with two key ingredients. First, DePT plugs visual prompts into the vision Transformer and only tunes these source-initialized prompts during adaptation. We find such parameter-efficient finetuning can efficiently adapt the model representation to the target domain without overfitting to the noise in the learning objective. Second, DePT bootstraps the source representation to the target domain by memory bank-based online pseudo labeling. A hierarchical self-supervised regularization specially designed for prompts is jointly optimized to alleviate error accumulation during self-training. With much fewer tunable parameters, DePT demonstrates not only state-of-the-art performance on major adaptation benchmarks, but also superior data efficiency, i.e., adaptation with only 1\% or 10\% data without much performance degradation compared to 100\% data. In addition, DePT is also versatile to be extended to online or multi-source TTA settings.

****

Mitigating Dataset Bias by Using Per-Sample Gradient

Sumyeong Ahn,Seongyoon Kim,Se-Young Yun

The performance of deep neural networks is strongly influenced by the training dataset setup. In particular, when attributes with a strong correlation with the target attribute are present, the trained model can provide unintended prejudgments and show significant inference errors (i.e., the dataset bias problem). Various methods have been proposed to mitigate dataset bias, and their emphasis is on weakly correlated samples, called bias-conflicting samples. These methods are based on explicit bias labels provided by humans. However, such methods require human costs. Recently, several studies have sought to reduce human intervention by utilizing the output space values of neural networks, such as feature space, logits, loss, or accuracy. However, these output space values may be insufficient for the model to understand the bias attributes well. In this study, we propose a debiasing algorithm leveraging gradient called Per-sample Gradient-based Debiasing (PGD). PGD is comprised of three steps: (1) training a model on uniform batch sampling, (2) setting the importance of each sample in proportion to the norm of the sample gradient, and (3) training the model using importance-batch sampling, whose probability is obtained in step (2). Compared with existing baselines for various datasets, the proposed method showed state-of-the-art accuracy for the classification task. Furthermore, we describe theoretical understandings of how PGD can mitigate dataset bias.

****

CAMA: A New Framework for Safe Multi-Agent Reinforcement Learning Using Constraint Augmentation

Ziyan Wang,Yali Du,Aivar Sootla,Haitham Bou Ammar,Jun Wang

With the widespread application of multi-agent reinforcement learning (MARL) in real-life settings, the ability to meet safety constraints has become an urgent problem to solve. For example, it is necessary to avoid collisions to reach a common goal in controlling multiple drones. We address this problem by introducing the Constraint Augmented Multi-Agent framework --- CAMA. CAMA can serve as a plug-and-play module to the popular MARL algorithms, including centralized training, decentralized execution and independent learning frameworks. In our approach, we represent the safety constraint as the sum of discounted safety costs bounded by the predefined value, which we call the safety budget. Experiments demonstrate that CAMA can converge quickly to a high degree of constraint satisfaction and surpasses other state-of-the-art safety counterpart algorithms in both cooperative and competitive settings.

****

CWATR: Generating Richer Captions with Object Attributes

Enes Muvahhid ■ahin,Gözde BOZDA■I AKAR

Image captioning is a popular yet challenging task which is at the intersection of Computer Vision and Natural Language Processing. Recently, transformer-based unified Vision and Language models advanced the state-of-the-art further on image captioning. However, there are still fundamental problems in these models. Even though the generated captions by these models are grammatically correct and describe the input image fairly good, they might overlook important details in the image. In this paper, we demonstrate these problems in a state-of-the-art baseline image captioning method and analyze the reasoning behind these problems. We propose a novel approach, named CWATR (Captioning With ATtRibutes), to integrate object attributes to the generated captions in order to obtain richer and more detailed captions. Our analyses demonstrate that the proposed approach generates richer and more visually grounded captions by integrating attributes of the objects in the scene to the generated captions successfully.

****

Internal Purity: A Differential Entropy based Internal Validation Index for Clustering Validation

Bin Cao,Chen Yang,Kaibo He,JING FAN

In an effective process of cluster analysis, it is indispensable to validate the goodness of different partitions after clustering. Existing internal validation indices are implemented based on distance, variance and model-selection. The indices based on distance or variance cannnot catpure the real ``density" of the cluster and the time complexity for distance based indices is usually too high to be applied for large datasets. Moreover, the indices based on model-selection tend to overestimate the number of cluster in clustering validation. Therefore, we propose a novel internal validation index based on the differential entropy, named \textit{internal purity} (IP). The proposed IP index can effectively measure the purity of a cluster without using the external cluster information, and successfully overcome the drawbacks of existing internal indices. Based on six powerful deep pre-trained models and without further fine-tuning using the experimental datasets, we use four different clustering algorithms to compare our index with thirteen other well-known internal indices on five text and five image datasets. The results show that, for 60 test cases in total, our IP index can return the optimal clustering results in 43 cases while the second best index can merely report the optimal partition in 17 cases, which demonstrates the significant superiority of our IP index when validating the goodness of the clustering results. Moreover, theoretical analysis for the effectiveness and efficiency of the proposed index are also provided.

****

Task Regularized Hybrid Knowledge Distillation For Continual Object Detection

Jinming Zhang,Mengxue Kang,Jinpeng Zhang,Zhe Ma,Shan Yu

Knowledge distillation has been used to overcome catastrophic forgetting in Continual Object Detection(COD) task. Previous works mainly focus on combining different distillation methods, including feature, classification, location and relation, into a mixed scheme to solve this problem. In this paper, we propose a task regularized hybrid knowledge distillation method for COD task. First, we propose an image-level hybrid knowledge representation by combining instance-level hard and soft knowledge to use teacher knowledge critically. Second, we propose a task-based regularization distillation loss by taking account of loss and category differences to make continual learning more balance between old and new tasks.

We find that, under appropriate knowledge selection and transfer strategies, using only classification distillation can also relieve knowledge forgetting effectively. Extensive experiments conducted on MS COCO2017 demonstrate that our method achieves state-of-the-art results under various scenarios. We get an absolute improvement of 27.98 at $RelGap$ under the most difficult five-task scenario. Code is in attachment and will be available on github.

****

Efficient Model Updates for Approximate Unlearning of Graph-Structured Data
Eli Chien,Chao Pan,Olgica Milenkovic
With the adoption of recent laws ensuring the ``right to be forgotten'', the problem of machine unlearning has become of significant importance. This is particularly the case for graph-structured data, and learning tools specialized for such data, including graph neural networks (GNNs). This work introduces the first known approach for \emph{approximate graph unlearning} with provable theoretical guarantees. The challenges in addressing the problem are two-fold. First, there exist multiple different types of unlearning requests that need to be considered, including node feature, edge and node unlearning. Second, to establish provable performance guarantees, one needs to carefully evaluate the process of feature mixing during propagation. We focus on analyzing Simple Graph Convolutions (SGC) and their generalized PageRank (GPR) extensions, thereby laying the theoretical foundations for unlearning GNNs. Empirical evaluations of six benchmark datasets demonstrate excellent performance/complexity/privacy trade-offs of our approach compared to complete retraining and general methods that do not leverage graph information. For example, unlearning 200 out of 1208 training nodes of the Cora dataset only leads to a 0.1% loss in test accuracy, but offers a 4-fold speed-up compared to complete retraining with a $(\epsilon,\delta)=(1,10^{-4})$ ``privacy cost''. We also exhibit a 12% increase in test accuracy for the same dataset when compared to unlearning methods that do not leverage graph information, with comparable time complexity and the same privacy guarantee.

****

Risk-aware Bayesian RL for Cautious Exploration
Rohan Mitta,Hosein Hasanbeig,Daniel Kroening,Alessandro Abate
This paper addresses the problem of maintaining safety during training in Reinforcement Learning (RL), such that the safety constraint violations are bounded at any point during learning. Whilst enforcing safety during training might limit the agent's exploration, we propose a new architecture that handles the trade-off between efficient progress in exploration and safety maintenance.
As the agent's exploration progresses, we update Dirichlet-Categorical models of the transition probabilities of the Markov decision process that describes the agent's behavior within the environment by means of Bayesian inference. We then propose a way to approximate moments of the agent's belief about the risk associated with the agent's behavior originating from local action selection. We demonstrate that this approach can be easily coupled with RL, we provide rigorous theoretical guarantees, and we present experimental results to showcase the performance of the overall architecture.

****

Change Detection for bi-temporal images classification based on Siamese Variational AutoEncoder and Transfer Learning
Chouikhi Farah,Ali Ben Abbes,Imed Riadh Farah
Siamese structures empower Deep Learning (DL) models to increase their efficiency by learning how to extract the relevant temporal features from the input data. In this paper, a Siamese Variational Auto-Encoder (VAE) model based on transfer learning (TL) is applied for change detection (CD) using bi-temporal images. The introduced method is trained in a supervised strategy for classification tasks. Firstly, the suggested generative method utilizes two VAEs to extract features from bi-temporal images. Subsequently, concatenates them into a feature vector. To get a classification map of the source scene, the classifier receives this vector and the ground truth data as input. The source model is fine-tuned to be applied to the target scene with less ground truth data using a TL strategy. Experiments were carried out in two study areas in the arid regions of southern Tun

isia. The obtained results reveal that the proposed method outperformed the Siamese Convolution Neural Network (SCNN) by achieving an accuracy of more than 98%, in the source scene, and increased the accuracy in the target scene by 1.25% by applying the TL strategy.

****

Learning Top-k Classification with Label Ranking

Bin Cao,Kai Wang,JING FAN,Jianwei Yin

Class confusability and multi-label nature of examples inevitably arise in classification tasks with the increasing number of classes, which poses a huge challenge to classification. To mitigate this problem, top-k classification is proposed, where the classifier is allowed to predict k label candidates and the prediction result is considered correct as long as the ground truth label is included in the k labels. However, existing top-k classification methods neglect the ranking of the ground truth label among the predicted k labels, which has high application value. In this paper, we propose a novel three-stage approach to learn top-k classification with label ranking. We first propose an ensemble based relabeling method and relabel the training data with k labels, which is used to train the top-k classifier. We then propose a novel top-k classification loss function that aims to improve the ranking of the ground truth label. Finally, we have conducted extensive experiments on four text datasets and four image datasets, and the experimental results show that our method could significantly improve the performance of existing methods.

****

Populating memory in Continual Learning with Consistency Aware Sampling

Julio Hurtado,Alain Raymond,Vladimir Araujo,Vincenzo Lomonaco,Alvaro Soto,Davide Bacciu

Continual Learning (CL) methods aim to mitigate Catastrophic Forgetting (CF), where knowledge from previously learned tasks is often lost in favor of the new one. Among those algorithms, some have shown the relevance of keeping a rehearsal buffer with previously seen examples, referred to as $memory$. Yet, despite their popularity, limited research has been done to understand which elements are more beneficial to store in memory. It is common for this memory to be populated through random sampling, with little guiding principles that may aid in retaining prior knowledge. In this paper, and consistent with previous work, we found that some storage policies behave similarly given a certain memory size or compute budget, but when these constraints are relevant, results differ considerably. Based on these insights, we propose CAWS (Consistency AWare Sampling), an original storage policy that leverages a learning consistency score (C-Score) to populate the memory with elements that are $easy$ to $learn$ and $representative$ of previous tasks. Because of the impracticality of directly using the C-Score in CL, we propose more feasible and efficient proxies to calculate the score that yield state-of-the-art results on CIFAR-100 and Tiny Imagenet.

****

A Unified Algebraic Perspective on Lipschitz Neural Networks

Alexandre Araujo,Aaron J Havens,Blaise Delattre,Alexandre Allauzen,Bin Hu

Important research efforts have focused on the design and training of neural networks with a controlled Lipschitz constant. The goal is to increase and sometimes guarantee the robustness against adversarial attacks. Recent promising techniques draw inspirations from different backgrounds to design 1-Lipschitz neural networks, just to name a few: convex potential layers derive from the discretization of continuous dynamical systems, Almost-Orthogonal-Layer proposes a tailored method for matrix rescaling. However, it is today important to consider the recent and promising contributions in the field under a common theoretical lens to better design new and improved layers. This paper introduces a novel algebraic perspective unifying various types of 1-Lipschitz neural networks, including the ones previously mentioned, along with methods based on orthogonality and spectral methods. Interestingly, we show that many existing techniques can be derived and generalized via finding analytical solutions of a common semidefinite programming (SDP) condition. We also prove that AOL biases the scaled weight to the ones which are close to the set of orthogonal matrices in a certain mathematical ma nner. Moreover, our algebraic condition, combined with the Gershgorin circle theorem, readily leads to new and diverse parameterizations for 1-Lipschitz network layers. Our approach, called SDP-based Lipschitz Layers (SLL), allows us to design non-trivial yet efficient generalization of convex potential layers. Finally, the comprehensive set of experiments on image classification shows that SLLs outperform previous approaches on certified robust accuracy. Code is available at https://github.com/araujoalexandre/Lipschitz-SLL-Networks.

****

AudioGen: Textually Guided Audio Generation
Felix Kreuk,Gabriel Synnaeve,Adam Polyak,Uriel Singer,Alexandre Défossez,Jade Copet,Devi Parikh,Yaniv Taigman,Yossi Adi

In this work, we tackle the problem of generating audio samples conditioned on descriptive text captions. We propose AudioGen, an auto-regressive generative model, operating on a learnt discrete audio representation, that generates audio samples conditioned on text inputs. The task of text-to-audio generation poses multiple challenges. Due to the way audio travels through a medium, differentiating ``objects'' can be a difficult task (e.g., separating multiple people simultaneously speaking). This is further complicated by real-world recording conditions (e.g., background noise, reverberation, etc.). Scarce text annotations impose another constraint, limiting the ability to scale models. Finally, modeling high fidelity audio requires one to operate over extremely long sequences. To alleviate the aforementioned challenges we propose an augmentation technique that mixes different audio samples, driving the model to internally learn to separate multiple sources. We curated 10 datasets containing different types of audio and text annotations to handle the scarcity of text-audio data points. For faster inference, we explore the use of multi-stream modeling, allowing the use of shorter sequences while maintaining a similar bitrate and perceptual quality. Finally, we apply classifier-free guidance to improve adherence to text. Comparing to the evaluated baselines, AudioGen outperforms over both objective and subjective metrics. We further conduct an ablation study to gauge the effects of pre-trained text and audio components.

****

Faster Reinforcement Learning with Value Target Lower Bounding
Le Zhao,Wei Xu

We show that an arbitrary lower bound of the maximum achievable value can be used to improve the Bellman value target during value learning. In the tabular case, value learning using the lower bounded Bellman operator converges to the same optimal value as using the original Bellman operator, at a potentially faster speed. In practice, discounted episodic return in episodic tasks and n-step bootstrapped return in continuing tasks can serve as lower bounds to improve the value target. We experiment on Atari games, FetchEnv tasks and a challenging physically simulated car push and reach task. We see large gains in sample efficiency as well as converged performance over common baselines such as TD3, SAC and Hindsight Experience Replay (HER) in most tasks, and observe a reliable and competitive performance against the stronger n-step methods such as td-lambda, Retrace and optimality tightening. Prior works have already successfully applied a special case of lower bounding (using episodic return), but are limited to a small number of episodic tasks. To the best of our knowledge, we are the first to propose the general method of value target lower bounding (with possibly bootstrapped return), to demonstrate its optimality in theory, and effectiveness in a wide range of tasks over many strong baselines.

****

Hebbian and Gradient-based Plasticity Enables Robust Memory and Rapid Learning in RNNs
Yu Duan,Zhongfan Jia,Qian Li,Yi Zhong,Kaisheng Ma

Rapidly learning from ongoing experiences and remembering past events with a flexible memory system are two core capacities of biological intelligence. While the underlying neural mechanisms are not fully understood, various evidence supports that synaptic plasticity plays a critical role in memory formation and fast learning. Inspired by these results, we equip Recurrent Neural Networks (RNNs) wi

th plasticity rules to enable them to adapt their parameters according to ongoing experiences. In addition to the traditional local Hebbian plasticity, we propose a global, gradient-based plasticity rule, which allows the model to evolve towards its self-determined target. Our models show promising results on sequential and associative memory tasks, illustrating their ability to robustly form and retain memories. In the meantime, these models can cope with many challenging few-shot learning problems. Comparing different plasticity rules under the same framework shows that Hebbian plasticity is well-suited for several memory and associative learning tasks; however, it is outperformed by gradient-based plasticity on few-shot regression tasks which require the model to infer the underlying mapping.
****

Towards Minimax Optimal Reward-free Reinforcement Learning in Linear MDPs
Pihe Hu,Yu Chen,Longbo Huang

We study reward-free reinforcement learning with linear function approximation for episodic Markov decision processes (MDPs). In this setting, an agent first interacts with the environment without accessing the reward function in the exploration phase. In the subsequent planning phase, it is given a reward function and asked to output an ϵ-optimal policy. We propose a novel algorithm LSVI-RFE under the linear MDP setting, where the transition probability and reward functions are linear in a feature mapping. We prove an $\widetilde{O}(H^{4} d^{2}/\epsilon^2)$ sample complexity upper bound for LSVI-RFE, where H is the episode length and d is the feature dimension. We also establish a sample complexity lower bound of $\Omega(H^{3} d^{2}/\epsilon^2)$. To the best of our knowledge, LSVI-RFE is the first computationally efficient algorithm that achieves the minimax optimal sample complexity in linear MDP settings up to an H and logarithmic factors. Our LSVI-RFE algorithm is based on a novel variance-aware exploration mechanism to avoid overly-conservative exploration in prior works. Our sharp bound relies on the decoupling of UCB bonuses during two phases, and a Bernstein-type self-normalized bound, which remove the extra dependency of sample complexity on H and d, respectively.
****

Context and History Aware Other-Shaping
Akbir Khan,Newton Kwan,Timon Willi,Chris Lu,Andrea Tacchetti,Jakob Nicolaus Foerster

Cooperation failures, in which self-interested agents converge to collectively worst-case outcomes, are a common failure mode of Multi-Agent Reinforcement Learning (MARL) methods. Methods such as Model-Free Opponent Shaping (MFOS) and The Good Shepherd address this issue by shaping their co-player’s learning into mutual cooperation. However, these methods fail to capture important co-player learning dynamics or do not scale to co-players parameterised by deep neural networks. To address these issues, we propose Context and History Aware Other-Shaping (CHAOS). A CHAOS agent is a meta-learner parameterised by a recurrent neural network that learns to shape its co-player over multiple trials. CHAOS considers both the context (inter-episode information), and history (intra-episode information) to shape co-players successfully. CHAOS also successfully scales to shaping co-players parameterised by deep neural networks. In a set of experiments, we show that CHAOS achieves state-of-the-art shaping in matrix games. We provide extensive ablations, motivating the importance of both context and history. CHAOS also successfully shapes on a complex grid-worldbased game, demonstrating CHAOS’s scalability empirically. Finally, we provide empirical evidence that, counterintuitively, the widely-used Coin Game environment does not require history to learn shaping because states are often indicative of past actions. This suggests that the Coin Game is, in contrast to common understanding, unsuitable for investigating shaping in high-dimensional, multi-step environments.
****

ReD-GCN: Revisit the Depth of Graph Convolutional Network
Yuchen Yan,Yuzhong Chen,Mahashweta Das,Hao Yang,Hanghang Tong

Finding the proper depth d of a GNN that provides strong representation power has drawn significant attention, yet nonetheless largely remains an open probl em for the graph learning community. Although noteworthy progress has been made, the depth or the number of layers of a corresponding GCN is realized by a series of graph convolution operations, which naturally makes d a positive integer ($d \in \mathbb{N}+$). An interesting question is whether breaking the constraint of $\mathbb{N}+$ by making d a real number ($d \in \mathbb{R}$) can bring new insights into graph learning mechanisms. In this work, by redefining GCN's depth d as a trainable parameter continuously adjustable within $(-\infty,+\infty)$, we open a new door of controlling its expressiveness on graph signal processing to model graph homophily/heterophily (nodes with similar/dissimilar labels/attributes tend to inter-connect). A simple and powerful GCN model ReD-GCN, is proposed to retain the simplicity of GCN and meanwhile automatically search for the optimal d without the prior knowledge regarding whether the input graph is homophilic or heterophilic. Negative-valued d intrinsically enables high-pass frequency filtering functionality for graph heterophily. Variants extending the model flexibility/scalability are also developed. The theoretical feasibility of having a real-valued depth with explainable physical meanings is ensured via eigen-decomposition of the graph Laplacian and a properly designed transformation function from the perspective of functional calculus. Extensive experiments demonstrate the superiority of ReD-GCN on node classification tasks for a variety of graphs. Furthermore, by introducing the concept of eigengraph, a novel graph augmentation method is obtained: the optimal d effectively generates a new topology through a properly weighted combination of eigengraphs, which dramatically boosts the performance even for a vanilla GCN.

****

The Influence of Learning Rule on Representation Dynamics in Wide Neural Networks

Blake Bordelon,Cengiz Pehlevan

It is unclear how changing the learning rule of a deep neural network alters its learning dynamics and representations. To gain insight into the relationship between learned features, function approximation, and the learning rule, we analyze infinite-width deep networks trained with gradient descent (GD) and biologically-plausible alternatives including feedback alignment (FA), direct feedback alignment (DFA), and error modulated Hebbian learning (Hebb), as well as gated linear networks (GLN). We show that, for each of these learning rules, the evolution of the output function at infinite width is governed by a time varying effective neural tangent kernel (eNTK). In the lazy training limit, this eNTK is static and does not evolve, while in the rich mean-field regime this kernel's evolution can be determined self-consistently with dynamical mean field theory (DMFT). This DMFT enables comparisons of the feature and prediction dynamics induced by each of these learning rules. In the lazy limit, we find that DFA and Hebb can only learn using the last layer features, while full FA can utilize earlier layers with a scale determined by the initial correlation between feedforward and feedback weight matrices. In the rich regime, DFA and FA utilize a temporally evolving and depth-dependent NTK. Counterintuitively, we find that FA networks trained in the rich regime exhibit more feature learning if initialized with smaller correlation between the forward and backward pass weights. GLNs admit a very simple formula for their lazy limit kernel and preserve conditional Gaussianity of their preactivations under gating functions. Error modulated Hebb rules show very small task-relevant alignment of their kernels and perform most task relevant learning in the last layer.

****

Multiple Modes for Continual Learning

Siddhartha Datta,Nigel Shadbolt

Adapting model parameters to incoming streams of data is a crucial factor to deep learning scalability. Interestingly, prior continual learning strategies in online settings inadvertently anchor their updated parameters to a local parameter subspace to remember old tasks, else drift away from the subspace and forget. From this observation, we formulate a trade-off between constructing multiple parameter modes and allocating tasks per mode. Mode-Optimized Task Allocation (MOTA), our contributed adaptation strategy, trains multiple modes in parallel, then

optimizes task allocation per mode. We empirically demonstrate improvements over baseline continual learning strategies and across varying distribution shifts, namely sub-population, domain, and task shift.
****
A Theory of Equivalence-Preserving Program Embeddings
Logan Weber,Jesse Michel,Alex Renda,Saman Amarasinghe,Michael Carbin
Program embeddings are used to solve tasks such as \textit{code clone detection} and \textit{semantic labeling}. Solutions to these \textit{semantic tasks} should be invariant to semantics-preserving program transformations. When a program embedding function satisfies this invariance, we call it an \textit{equivalence-preserving program embedding function}. We say a programming language can be \textit{tractably embedded} when we can construct an equivalence-preserving program embedding function that executes in polynomial time in program/input length and produces program embeddings that are proportional to the input length. Determining whether a programming language can be tractably embedded is the \textit{equivalence-preserving program embedding problem}. We formalize this problem and theoretically characterize when programming languages can be tractably embedded. To validate our theoretical results, we use the BERT-Tiny model to learn an equivalence-preserving program embedding function for a programming language that can be tractably embedded and show the model fails to construct an equivalence-preserving program embedding function for a similar language that is intractable to embed.

****
On the Data-Efficiency with Contrastive Image Transformation in Reinforcement Learning
Sicong Liu,Xi Sheryl Zhang,Yushuo Li,Yifan Zhang,Jian Cheng
Data-efficiency has always been an essential issue in pixel-based reinforcement learning (RL). As the agent not only learns decision-making but also meaningful representations from images. The line of reinforcement learning with data augmentation shows significant improvements in sample-efficiency. However, it is challenging to guarantee the optimality invariant transformation, that is, the augmented data are readily recognized as a completely different state by the agent. In the end, we propose a contrastive invariant transformation (CoIT), a simple yet promising learnable data augmentation combined with standard model-free algorithms to improve sample-efficiency. Concretely, the differentiable CoIT leverages original samples with augmented samples and hastens the state encoder for a contrastive invariant embedding. We evaluate our approach on DeepMind Control Suite and Atari100K. Empirical results verify advances using CoIT, enabling it to outperform the new state-of-the-art on various tasks. Source code is available at https://github.com/mooricAnna/CoIT.
****
Energy-based Out-of-Distribution Detection for Graph Neural Networks
Qitian Wu,Yiting Chen,Chenxiao Yang,Junchi Yan
Representation learning on semi-structured data, e.g., graphs, has become a central problem in deep learning community as relational structures are pervasive in real situations and induce data inter-dependence that hinders trivial adaptation of existing approaches in other domains where the inputs are assumed to be i.i.d. sampled. However, current models in this regime mostly focus on improving testing performance of in-distribution data and largely ignores the potential risk w.r.t. out-of-distribution (OOD) testing samples that may cause negative outcome if the model is overconfident in prediction on them. In this paper, we identify a provably effective OOD discriminator based on an energy function directly extracted from a graph neural network trained with standard supervised classification loss. This paves a way for a simple and efficient OOD detection model for GNN-based semi-supervised learning on graphs, which we call GNN-Safe. It also has nice theoretical properties that guarantee an overall distinguishable margin between the detection scores for in-distribution and OOD samples, which, more critically, can be further strengthened by a non-learning-based structured propagation scheme. Extensive experiments over five real-world datasets validate the pract

ical efficacy of the proposed model for detecting various OOD instances that are inter-connected in a graph with up to 17.0% improvement on average AUROC over competitive peer models and without sacrificing in-distribution testing accuracy.

****

Theoretical Characterization of Neural Network Generalization with Group Imbalance

Hongkang Li,Shuai Zhang,Meng Wang,Yihua Zhang,Pin-Yu Chen,Sijia Liu

Group imbalance has been a known problem in empirical risk minimization (ERM), where the achieved high \textit{average} accuracy could be accompanied by low accuracy in a \textit{minority} group. Despite various algorithmic efforts to improve the minority group accuracy, a theoretical study of the generalization performance of ERM on individual groups remains elusive. By formulating the group imbalance problem with the Gaussian Mixture Model, this paper quantifies the impact of individual groups on the sample complexity, the convergence rate, and the average and group-level testing performance. Although our theoretical framework is centered on binary classification using a one-hidden-layer neural network, to the best of our knowledge, we provide the first theoretical analysis of the group-level generalization of ERM in addition to the commonly studied average generalization performance. Sample insights of our theoretical results include that when all group-level co-variance is in the medium regime and all mean are close to zero, the learning performance is most desirable in the sense of a small sample complexity, a fast training rate, and a high average and group-level testing accuracy. Moreover, we show that increasing the fraction of the minority group in the training data does not necessarily improve the generalization performance of the minority group. Our theoretical results are validated on both synthetic and empirical datasets such as CelebA and CIFAR-10 in image classification.

****

Formal Interpretability with Merlin-Arthur Classifiers

Stephan Waeldchen,Kartikey Sharma,Max Zimmer,Berkant Turan,Sebastian Pokutta

We propose a new type of multi-agent interactive classifier that provides, for the first time, provable interpretability guarantees even for complex agents such as neural networks.   In our setting, which is inspired by the Merlin-Arthur protocol from Interactive Proof Systems, two agents cooperate to provide a classification: the prover selects a small set of features as a certificate and presents it to the verifier who decides the class. A second, adversarial prover ensures the truthfulness of the system and allows us to connect the game-theoretic equilibrium between the provers and the verifier to guarantees on the exchanged features. We define completeness and soundness metrics to provide a lower bound on the mutual information between the features and the class. Our experiments demonstrate good agreement between theory and practice using neural network classifiers, and we show how our setup practically prevents manipulation.

****

Quasi-optimal Reinforcement Learning with Continuous Actions

Yuhan Li,Wenzhuo Zhou,Ruoqing Zhu

Many real-world applications of reinforcement learning (RL) require making decisions in continuous action environments. In particular, determining the optimal dose level plays a vital role in developing medical treatment regimes. One challenge in adapting existing RL algorithms to medical applications, however, is that the popular infinite support stochastic policies, e.g., Gaussian policy, may assign riskily high dosages and harm patients seriously. Hence, it is important to induce a policy class whose support only contains near-optimal actions, and shrink the action-searching area for effectiveness and reliability. To achieve this, we develop a novel quasi-optimal learning algorithm, which can be easily optimized in off-policy settings with guaranteed convergence under general function approximations. Theoretically, we analyze the consistency, sample complexity, adaptability, and convergence of the proposed algorithm. We evaluate our algorithm with comprehensive simulated experiments and a dose suggestion real application to Ohio Type 1 diabetes dataset.

****

Generalization Bounds for Federated Learning: Fast Rates, Unparticipating Client s and Unbounded Losses

Xiaolin Hu,Shaojie Li,Yong Liu

In {federated learning}, the underlying data distributions may be different across clients. This paper provides a theoretical analysis of generalization error of {federated learning}, which captures both heterogeneity and relatedness of the distributions. In particular, we assume that the heterogeneous distributions are sampled from a meta-distribution. In this two-level distribution framework, we characterize the generalization error not only for clients participating in the training but also for unparticipating clients. We first show that the generalization error for unparticipating clients can be bounded by participating generalization error and participating gap caused by clients' sampling. We further establish fast learning bounds of order $\mathcal{O}(\frac{1}{mn} + \frac{1}{m})$ for unparticipating clients, where m is the number of clients and n is the sample size at each client. To our knowledge, the obtained fast bounds are state-of-the-art in the two-level distribution framework. Moreover, previous theoretical results mostly require the loss function to be bounded. We derive convergence bounds of order $\mathcal{O}(\frac{1}{\sqrt{mn}} + \frac{1}{\sqrt{m}})$ under unbounded assumptions, including sub-exponential and sub-Weibull losses.

****

Contrastive Unsupervised Learning of World Model with Invariant Causal Features

Rudra P. K. Poudel,Harit Pandya,Roberto Cipolla

In this paper we present a world model, which learns the causal features using invariance principle. We use contrastive unsupervised learning to learn the invariant causal features, which enforces invariance across augmentations of irrelevant parts or styles of the observation. Since the world model based reinforcement learning methods optimize representation learning and policy of the agent independently, contrastive loss collapses due to lack of supervisory signal to the representation learning module. We propose depth reconstruction as an auxiliary task to explicitly enforce the invariance and data augmentation as style intervention on the RGB space to mitigate this issue. Our design help us to leverage state-of-the-art unsupervised representation learning method to learn the world model with invariant causal features, which outperforms current state-of-the-art model-based as well as model-free reinforcement learning methods on out-of-distribution point navigation tasks on Gibson and iGibson dataset at 100k and 500k interaction step benchmarks. Further experiments on DeepMind control suite even without depth reconstruction, our proposed model performs on par with the state-of-the-art counterpart models.

****

GOING BEYOND 1-WL EXPRESSIVE POWER WITH 1-LAYER GRAPH NEURAL NETWORKS

Tianjun Yao,Yingxu Wang,Shangsong Liang

Graph neural networks have become the \textit{de facto} standard for representational learning in graphs, and have achieved SOTA in many graph-related tasks such as node classification, graph classification and link prediction. However, it has been shown that the expressive power is equivalent maximally to Weisfeiler-Lehman Test. Recently, there is a line of work aiming to enhance the expressive power of graph neural networks. In this work, we propose a more generalized variant of neural Weisfeiler-Lehman test to enhance structural representation for each node in a graph to uplift the expressive power of any graph neural network. It is shown theoretically our method is strictly more powerful than 1\&2-WL test. The Numerical experiments also demonstrate that our proposed method outperforms the standard GNNs on almost all the benchmark datasets by a large margin in most cases with significantly lower running time and memory consumption compared with other more powerful GNNs.

****

System Identification as a Reinforcement Learning Problem

Jose Antonio Martin H.,Oscar Fernández Vicente,Sergio Perez,Anas Belfadil,Cristina Ibanez-Llano,Freddy Perozo,Jose Javier Valle,Javier Arechalde Pelaz

System identification, also known as learning forward models, transfer functions, system dynamics, etc., has a long tradition both in science and engineering in different fields. Particularly, it is a recurring theme in Reinforcement Learni ng research, where forward models approximate the state transition function of a Markov Decision Process by learning a mapping function from current state and action to the next state. This problem is commonly defined as a Supervised Learning problem in a direct way. This common approach faces several difficulties due to the inherent complexities of the dynamics to learn, for example, delayed effects, high non-linearity, non-stationarity, partial observability and, more important, error accumulation when using bootstrapped predictions (predictions based on past predictions), over large time horizons. Here we explore the use of Reinforcement Learning in this problem. We elaborate on why and how this problem fits naturally and sound as a Reinforcement Learning problem, and present some experimental results that demonstrate RL is a promising technique to solve these kind of problems.

****

When to Trust Aggregated Gradients: Addressing Negative Client Sampling in Federated Learning

Wenkai Yang,Yankai Lin,Guangxiang Zhao,Peng Li,Jie Zhou,Xu Sun

Federated Learning has become a widely-used framework which allows learning a global model on decentralized local datasets under the condition of protecting local data privacy. However, federated learning faces severe optimization difficulty when training samples are not independently and identically distributed (non-i.i.d.). In this paper, we point out that the client sampling practice plays a decisive role in the aforementioned optimization difficulty. We find that the negative client sampling will cause the merged data distribution of currently sampled clients heavily inconsistent with that of all available clients, and further make the aggregated gradient unreliable. To address this issue, we propose a novel learning rate adaptation mechanism to adaptively adjust the server learning rate for the aggregated gradient in each round, according to the consistency between the merged data distribution of currently sampled clients and that of all available clients. Specifically, we make theoretical deductions to find a meaningful and robust indicator that is positively related to the optimal server learning rate and can effectively reflect the merged data distribution of sampled clients, and we utilize it for the server learning rate adaptation. Extensive experiments on multiple image and text classification tasks validate the great effectiveness of our method.

****

More ConvNets in the 2020s: Scaling up Kernels Beyond 51x51 using Sparsity

Shiwei Liu,Tianlong Chen,Xiaohan Chen,Xuxi Chen,Qiao Xiao,Boqian Wu,Tommi Kärkkäinen,Mykola Pechenizkiy,Decebal Constantin Mocanu,Zhangyang Wang

Transformers have quickly shined in the computer vision world since the emergence of Vision Transformers (ViTs). The dominant role of convolutional neural networks (CNNs) seems to be challenged by increasingly effective transformer-based models. Very recently, a couple of advanced convolutional models strike back with large kernels motivated by the local-window attention mechanism, showing appealing performance and efficiency. While one of them, i.e. RepLKNet, impressively manages to scale the kernel size to 31x31 with improved performance, the performance starts to saturate as the kernel size continues growing, compared to the scaling trend of advanced ViTs such as Swin Transformer. In this paper, we explore the possibility of training extreme convolutions larger than 31x31 and test whether the performance gap can be eliminated by strategically enlarging convolutions. This study ends up with a recipe for applying extremely large kernels from the perspective of sparsity, which can smoothly scale up kernels to 61x61 with better performance. Built on this recipe, we propose Sparse Large Kernel Network (SLaK), a pure CNN architecture equipped with sparse factorized 51x51 kernels that can perform on par with or better than state-of-the-art hierarchical Transformers and modern ConvNet architectures like ConvNeXt and RepLKNet, on ImageNet classification as well as a wide range of downstream tasks including semantic segmentation on ADE20K, object detection on PASCAL VOC 2007, and object detection/segmentation on MS COCO. Codes are available at https://github.com/VITA-Group/SLaK.

****

Few-shot Cross-domain Image Generation via Inference-time Latent-code Learning

Arnab Kumar Mondal,Piyush Tiwary,Parag Singla,Prathosh AP

In this work, our objective is to adapt a Deep generative model trained on a large-scale source dataset to multiple target domains with scarce data. Specifically, we focus on adapting a pre-trained Generative Adversarial Network (GAN) to a target domain without re-training the generator. Our method draws the motivation from the fact that out-of-distribution samples can be `embedded' onto the latent space of a pre-trained source-GAN. We propose to train a small latent-generation network during the inference stage, each time a batch of target samples is to be generated. These target latent codes are fed to the source-generator to obtain novel target samples. Despite using the same small set of target samples and the source generator, multiple independent training episodes of the latent-generation network results in the diversity of the generated target samples. Our method, albeit simple, can be used to generate data from multiple target distributions using a generator trained on a single source distribution. We demonstrate the efficacy of our surprisingly simple method in generating multiple target datasets with only a single source generator and a few target samples.

****

RLx2: Training a Sparse Deep Reinforcement Learning Model from Scratch

Yiqin Tan,Pihe Hu,Ling Pan,Jiatai Huang,Longbo Huang

Training deep reinforcement learning (DRL) models usually requires high computation costs. Therefore, compressing DRL models possesses immense potential for training acceleration and model deployment. However, existing methods that generate small models mainly adopt the knowledge distillation-based approach by iteratively training a dense network. As a result, the training process still demands massive computing resources. Indeed, sparse training from scratch in DRL has not been well explored and is particularly challenging due to non-stationarity in bootstrap training. In this work, we propose a novel sparse DRL training framework, “the Rigged Reinforcement Learning Lottery” (RLx2), which builds upon gradient-based topology evolution and is capable of training a sparse DRL model based entirely on a sparse network. Specifically, RLx2 introduces a novel multi-step TD target mechanism with a dynamic-capacity replay buffer to achieve robust value learning and efficient topology exploration in sparse models. It also reaches state-of-the-art sparse training performance in several tasks, showing $7.5\times$-$20\times$ model compression with less than 3% performance degradation and up to $20\times$ and $50\times$ FLOPs reduction for training and inference, respectively.

****

Rethinking Positive Sampling for Contrastive Learning with Kernel

Benoit Dufumier,Carlo Alberto Barbano,Robin Louiset,Edouard Duchesnay,Pietro Gori

Data augmentation is a crucial component in unsupervised contrastive learning (CL). It determines how positive samples are defined and, ultimately, the quality of the representation. Even if efforts have been made to find efficient augmentations for ImageNet, CL underperforms compared to supervised methods and it is still an open problem in other applications, such as medical imaging, or in datasets with easy-to-learn but irrelevant imaging features. In this work, we propose a new way to define positive samples using kernel theory along with a novel loss called \textit{decoupled uniformity}. We propose to integrate prior information, learnt from generative models viewed as feature extractor, or given as auxiliary attributes, into contrastive learning, to make it less dependent on data augmentation. We draw a connection between contrastive learning and the conditional mean embedding theory to derive tight bounds on the downstream classification loss. In an unsupervised setting, we empirically demonstrate that CL benefits from generative models, such as VAE and GAN, to less rely on data augmentations. We validate our framework on vision and medical datasets including CIFAR10, CIFAR100, STL10, ImageNet100, CheXpert and a brain MRI dataset. In the weakly supervised setting, we demonstrate that our formulation provides state-of-the-art results.

****

Stationary Deep Reinforcement Learning with Quantum K-spin Hamiltonian Equation

Xiao-Yang Liu,Zechu Li,Shixun Wu,Xiaodong Wang

Instability is a major issue of deep reinforcement learning (DRL) algorithms --- high variance of cumulative rewards over multiple runs. The instability is mainly caused by the existence of \textit{many local minimas} and worsened by the \textit{multiple fixed points} issue of Bellman's optimality equation. As a fix, we propose a quantum K-spin Hamiltonian regularization term (called \textit{H-term}) to help a policy network converge to a high-quality local minima. First, we take a quantum perspective by modeling a policy as a \textit{K-spin Ising model} and employ a Hamiltonian equation to measure the \textit{energy} of a policy. Then, we derive a novel Hamiltonian policy gradient theorem and design a generic actor-critic algorithm that utilizes the H-term to regularize the policy network. Finally, the proposed method significantly reduces the variance of cumulative rewards by $65.2\% \sim 85.6\%$ on six MuJoCo tasks; achieves an approximation ratio ≤ 1.05 over 90% test cases and reduces its variance by $60.16\% \sim 94.52\%$ on two combinatorial optimization tasks and two non-convex optimization tasks, compared with those of existing algorithms over 20 runs, respectively.

****

Performance Disparities Between Accents in Automatic Speech Recognition

Alex DiChristofano,Henry Shuster,Shefali Chandra,Neal Patwari

Automatic speech recognition (ASR) services are ubiquitous, transforming speech into text for systems like Amazon’s Alexa, Google’s Assistant, and Microsoft’s Cortana. Past research has identified discriminatory ASR performance as a function of racial group and nationality. In this paper, we expand the discussion about nationality and English language ASR by performing an audit of some of the most popular English ASR services using a large and global data set of speech from The Speech Accent Archive. We show that performance disparities exist as a function of whether or not a speaker’s first language is English and, even when controlling for multiple linguistic covariates, that these disparities have a statistically significant relationship to the political alignment of the speaker’s birth country with respect to the United States’ geopolitical power. We discuss this bias in the context of the historical use of language to maintain global and political power.

****

Do Perceptually Aligned Gradients Imply Robustness?

Roy Ganz,Bahjat Kawar,Michael Elad

Deep learning-based networks have achieved unprecedented success in numerous tasks, among which image classification. Despite these remarkable achievements, recent studies have demonstrated that such classification networks are easily fooled by small malicious perturbations, also known as adversarial examples. This security weakness led to extensive research aimed at obtaining robust models. Beyond the clear robustness benefits of such models, it was also observed that their gradients with respect to the input align with human perception. Several works have identified Perceptually Aligned Gradients (PAG) as a byproduct of robust training, but none have considered it as a standalone phenomenon nor studied its own implications. In this work, we focus on this trait and test whether \emph{Perceptually Aligned Gradients imply Robustness}. To this end, we develop a novel objective to directly promote PAG in training classifiers and examine whether models with such gradients are more robust to adversarial attacks. We present both heuristic and principled ways for obtaining target PAGs, which our method aims to learn. Specifically, we harness recent findings in score-based generative modeling as a source for PAG. Extensive experiments on CIFAR-10 and STL validate that models trained with our method have improved robust performance, exposing the surprising bidirectional connection between PAG and robustness.

****

Sparsity May Cry: Let Us Fail (Current) Sparse Neural Networks Together!

Shiwei Liu,Tianlong Chen,Zhenyu Zhang,Xuxi Chen,Tianjin Huang,AJAY KUMAR JAISWAL,Zhangyang Wang

Sparse Neural Networks (SNNs) have received voluminous attention predominantly due to growing computational and memory footprints of consistently exploding para meter count in large-scale models. Similar to their dense counterparts, recent SNNs generalize just as well and are equipped with numerous favorable benefits (e.g., low complexity, high scalability, and robustness), sometimes even better than the original dense networks. As research effort is focused on developing increasingly sophisticated sparse algorithms, it is startling that a comprehensive benchmark to evaluate the effectiveness of these algorithms has been highly overlooked. In absence of a carefully crafted evaluation benchmark, most if not all, sparse algorithms are evaluated against fairly simple and naive tasks (eg. CIFAR-10/100, ImageNet, GLUE, etc.), which can potentially camouflage many advantages as well unexpected predicaments of SNNs. In pursuit of a more general evaluation and unveiling the true potential of sparse algorithms, we introduce “Sparsity May Cry” Benchmark (SMC-Bench), a collection of carefully-curated 4 diverse tasks with 10 datasets, that accounts for capturing a wide range of domain-specific and sophisticated knowledge. Our systemic evaluation of the most representative sparse algorithms reveals an important obscured observation: the state-of-the-art magnitude- and/or gradient-based sparse algorithms seemingly fail to perform on SMC-Bench when applied out-of-the-box, sometimes at significantly trivial sparsity as low as 5%. The observations seek the immediate attention of the sparsity research community to reconsider the highly proclaimed benefits of SNNs. We further conduct a thorough investigation into the reasons for the failure of common SNNs. Our analysis points out that such failure is intimately related to the “lazy regime” of large model training, which hints us with stronger pruning recipes that alleviate the failure on SMC-Bench (though still more or less suffering). By incorporating these well-thought and diverse tasks, SMC-Bench is designed to favor and encourage the development of more scalable and generalizable sparse algorithms. We open-source SMC-Bench to assist researchers in building next-generation sparse algorithms that scale and generalize: https://github.com/VITA-Group/SMC-Bench.

****

Multitask Reinforcement Learning by Optimizing Neural Pathways

Samin Yeasar Arnob,Riyasat Ohib,Amy Zhang,Sergey Plis,Doina Precup

Reinforcement learning (RL) algorithms have achieved great success in learning specific tasks, as evidenced by examples such as AlphaGo or fusion control. However, it is still difficult for an RL agent to learn how to solve multiple tasks. In this paper, we propose a novel multitask learning framework, in which multiple specialized pathways through a single network are trained simultaneously, with each pathway focusing on a single task. We show that this approach achieves competitive performance with existing multitask RL methods, while using only 5% of the number of neurons per task. We demonstrate empirically the success of our approach on several continuous control tasks, in both online and offline training.

****

How deep convolutional neural networks lose spatial information with training

Umberto Maria Tomasini,Leonardo Petrini,Francesco Cagnetta,Matthieu Wyart

A central question of machine learning is how deep nets manage to learn tasks in high dimensions. An appealing hypothesis is that they achieve this feat by building a representation of the data where information irrelevant to the task is lost. For image data sets, this view is supported by the observation that after (and not before) training, the neural representation becomes less and less sensitive to diffeomorphisms acting on images as the signal propagates through the net.

This loss of sensitivity correlates with performance, and surprisingly also correlates with a gain of sensitivity to white noise acquired during training. These facts are unexplained, and as we demonstrate still hold when white noise is added to the images of the training set. Here, we (i) show empirically for various architectures that stability to image diffeomorphisms is achieved by spatial pooling in the first half of the net, and by channel pooling in the second half, (ii) introduce a scale-detection task for a simple model of data where pooling is learnt during training, which captures all empirical observations above and (iii) compute in this model how stability to diffeomorphisms and noise scale with depth. The scalings are found to depend on the presence of strides in the net ar chitecture. We find that the increased sensitivity to noise is due to the perturbing noise piling up during pooling, after a ReLU non-linearity is applied to the noise in the internal layers.

****

Which Layer is Learning Faster? A Systematic Exploration of Layer-wise Convergence Rate for Deep Neural Networks

Yixiong Chen,Alan Yuille,Zongwei Zhou

The deeply hierarchical structures enable deep neural networks (DNNs) to fit extremely complex target functions. However, the complex interaction between layers also makes the learning process of a particular layer poorly understood. This work demonstrates that the shallower layers of DNNs tend to converge faster than the deeper layers. We call this phenomenon Layer Convergence Bias. We also uncover the fundamental reason behind this phenomenon: Flatter local minima of shallower layers make their gradients more stable and predictive, allowing for faster training. Another surprising result is that the shallower layers tend to learn the low-frequency components of the target function, while the deeper layers usually learn the high-frequency components. It is consistent with the recent discovery that DNNs learn lower frequency objects faster.

****

Joint Embedding Self-Supervised Learning in the Kernel Regime

Bobak Kiani,Randall Balestriero,Yubei Chen,Seth Lloyd,Yann LeCun

The fundamental goal of self-supervised learning (SSL) is to produce useful representations of data without access to any labels for classifying the data. Modern methods in SSL, which form representations based on known or constructed relationships between samples, have been particularly effective at this task. Here, we aim to extend this framework to incorporate algorithms based on kernel methods where embeddings are constructed by linear maps acting on the feature space of a kernel. In this kernel regime, we derive methods to find the optimal form of the output representations for contrastive and non-contrastive loss functions. This procedure produces a new representation space with an inner product denoted as the induced kernel which generally correlates points which are related by an augmentation in kernel space and de-correlates points otherwise. We analyze our kernel model on small datasets to identify common features of self-supervised learning algorithms and gain theoretical insights into their performance on downstream tasks.

****

Linear convergence for natural policy gradient with log-linear policy parametrization

Carlo Alfano,Patrick Rebeschini

We analyze the convergence rate of the \emph{unregularized} natural policy gradient algorithm with log-linear policy parametrizations in infinite-horizon discounted Markov decision processes. In the deterministic case, when the Q-value is known and can be approximated by a linear combination of a known feature function up to a bias error, we show that a geometrically-increasing step size yields a linear convergence rate towards an optimal policy. We then consider the sample-based case, when the best representation of the Q-value function among linear combinations of a known feature function is known up to an estimation error. In this setting, we show that the algorithm enjoys the same linear guarantees as in the deterministic case up to an error term that depends on the estimation error, the bias error, and the condition number of the feature covariance matrix. Our results build upon the general framework of policy mirror descent and extend previous findings for the softmax tabular parametrization to the log-linear policy class.

****

A Non-Asymptotic Analysis of Oversmoothing in Graph Neural Networks

Xinyi Wu,Zhengdao Chen,William Wei Wang,Ali Jadbabaie

Oversmoothing is a central challenge of building more powerful Graph Neural Networks (GNNs). While previous works have only demonstrated that oversmoothing is inevitable when the number of graph convolutions tends to infinity, in this paper, we precisely characterize the mechanism behind the phenomenon via a non-asympt otic analysis. Specifically, we distinguish between two different effects when applying graph convolutions—an undesirable mixing effect that homogenizes node representations in different classes, and a desirable denoising effect that homogenizes node representations in the same class. By quantifying these two effects on random graphs sampled from the Contextual Stochastic Block Model (CSBM), we show that oversmoothing happens once the mixing effect starts to dominate the denoising effect, and the number of layers required for this transition is $O(\log N/\log (\log N))$ for sufficiently dense graphs with N nodes. We also extend our analysis to study the effects of Personalized PageRank (PPR), or equivalently, the effects of initial residual connections on oversmoothing. Our results suggest that while PPR mitigates oversmoothing at deeper layers, PPR-based architectures still achieve their best performance at a shallow depth and are outperformed by the graph convolution approach on certain graphs. Finally, we support our theoretical results with numerical experiments, which further suggest that the oversmoothing phenomenon observed in practice can be magnified by the difficulty of optimizing deep GNN models.

****

Class-Incremental Learning with Repetition

Hamed Hemati,Andrea Cossu,Antonio Carta,Julio Hurtado,Lorenzo Pellegrini,Davide Bacciu,Vincenzo Lomonaco,Damian Borth

Real-world data streams naturally include the repetition of previous concepts. From a Continual Learning (CL) perspective, repetition is a property of the environment and, unlike replay, cannot be controlled by the user. Nowadays, Class-Incremental scenarios represent the leading test-bed for assessing and comparing CL strategies. This family of scenarios is very easy to use, but it never allows revisiting previously seen classes, thus completely disregarding the role of repetition. We focus on the family of Class-Incremental with Repetition (CIR) scenarios, where repetition is embedded in the definition of the stream. We propose two stochastic scenario generators that produce a wide range of CIR scenarios starting from a single dataset and a few control parameters. We conduct the first comprehensive evaluation of repetition in CL by studying the behavior of existing CL strategies under different CIR scenarios. We then present a novel replay strategy that exploits repetition and counteracts the natural imbalance present in the stream. On both CIFAR100 and TinyImageNet, our strategy outperforms other replay approaches, which are not designed for environments with repetition.

****

Scaleformer: Iterative Multi-scale Refining Transformers for Time Series Forecasting

Mohammad Amin Shabani,Amir H. Abdi,Lili Meng,Tristan Sylvain

The performance of time series forecasting has recently been greatly improved by the introduction of transformers. In this paper, we propose a general multi-scale framework that can be applied to state-of-the-art transformer-based time series forecasting models
(FEDformer, Autoformer, etc.). Using iteratively refining a forecasted time series at multiple scales with shared weights, architecture adaptations and a specially-designed normalization scheme, we are able to achieve significant performance improvements with minimal additional computational overhead. Via detailed ablation studies, we demonstrate the effectiveness of our proposed architectural and methodological innovations. Furthermore, our experiments on various public datasets demonstrate that the proposed method outperforms the corresponding baselines. Depending on the choice of transformer architecture, our mutli-scale framework results in mean squared error reductions ranging from 5.5% to 38.5%. Our code is publicly available in https://github.com/BorealisAI/scaleformer.

****

Theoretical Characterization of How Neural Network Pruning Affects its Generalization

Hongru Yang,Yingbin Liang,Xiaojie Guo,Lingfei Wu,Zhangyang Wang

It has been observed in practice that applying pruning-at-initialization methods to neural networks and training the sparsified networks can not only retain the testing performance of the original dense models, but also sometimes even sligh tly boost the generalization performance. Theoretical understanding for such experimental observations are yet to be developed. This work makes the first attempt to study how different pruning fractions affect the model's gradient descent dynamics and generalization. Specifically, this work considers a classification task for overparameterized two-layer neural networks, where the network is randomly pruned according to different rates at the initialization. It is shown that as long as the pruning fraction is below a certain threshold, gradient descent can drive the training loss toward zero and the network exhibits good generalization performance. More surprisingly, the generalization bound gets better as the pruning fraction gets larger. To complement this positive result, this work further shows a negative result: there exists a large pruning fraction such that while gradient descent is still able to drive the training loss toward zero (by memorizing noise), the generalization performance is no better than random guessing. This further suggests that pruning can change the feature learning process, which leads to the performance drop of the pruned neural network. Up to our knowledge, this is the first generalization result for pruned neural networks, suggesting that pruning can improve the neural network's generalization.

****

Backdoors Stuck At The Frontdoor: Multi-Agent Backdoor Attacks That Backfire
Siddhartha Datta,Nigel Shadbolt

Malicious agents in collaborative learning and outsourced data collection threaten the training of clean models. Backdoor attacks, where an attacker poisons a model during training to successfully achieve targeted misclassification, are a major concern to train-time robustness. In this paper, we investigate a multi-agent backdoor attack scenario, where multiple attackers attempt to backdoor a victim model simultaneously. A consistent backfiring phenomenon is observed across a wide range of games, where agents suffer from a low collective attack success rate. We examine different modes of backdoor attack configurations, non-cooperation / cooperation, joint distribution shifts, and game setups to return an equilibrium attack success rate at the lower bound. The results motivate the re-evaluation of backdoor defense research for practical environments.

****

Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
Albert Qiaochu Jiang,Sean Welleck,Jin Peng Zhou,Timothee Lacroix,Jiacheng Liu,Wenda Li,Mateja Jamnik,Guillaume Lample,Yuhuai Wu

The formalization of existing mathematical proofs is a notoriously difficult process. Despite decades of research on automation and proof assistants, writing formal proofs remains arduous and only accessible to a few experts. While previous studies to automate formalization focused on powerful search algorithms, no attempts were made to take advantage of available informal proofs. In this work, we introduce Draft, Sketch, and Prove (DSP), a method that maps informal proofs to formal proof sketches, and uses the sketches to guide an automated prover by directing its search to easier sub-problems. We investigate two relevant setups where informal proofs are either written by humans or generated by a language model. Our experiments and ablation studies show that large language models are able to produce well-structured formal sketches that follow the same reasoning steps as the informal proofs. Guiding an automated prover with these sketches enhances its performance from 20.9% to 39.3% on a collection of mathematical competition problems.

****

Interpolating Compressed Parameter Subspaces
Siddhartha Datta,Nigel Shadbolt

Though distribution shifts have caused growing concern for machine learning scalability, solutions tend to specialize towards a specific type of distribution shift. Methods for label shift may not succeed against domain or task shift, and vice versa. We learn that constructing a Compressed Parameter Subspaces (CPS), a geometric structure representing distance-regularized parameters mapped to a set of train-time distributions, can maximize average accuracy over a broad range of distribution shifts concurrently. We show sampling parameters within a CPS can mitigate backdoor, adversarial, permutation, stylization and rotation perturbat ions. We also show training a hypernetwork representing a CPS can adapt to seen tasks as well as unseen interpolated tasks.

****

Liquid Structural State-Space Models

Ramin Hasani,Mathias Lechner,Tsun-Hsuan Wang,Makram Chahine,Alexander Amini,Daniela Rus

A proper parametrization of state transition matrices of linear state-space models (SSMs) followed by standard nonlinearities enables them to efficiently learn representations from sequential data, establishing the state-of-the-art on an extensive series of long-range sequence modeling benchmarks. In this paper, we show that we can improve further when the structured SSM, such as S4, is given by a linear liquid time-constant (LTC) state-space model. LTC neural networks are causal continuous-time neural networks with an input-dependent state transition module, which makes them learn to adapt to incoming inputs at inference. We show that by using a diagonal plus low-rank decomposition of the state transition matrix introduced in S4, and a few simplifications, the LTC-based structured state-space model, dubbed Liquid-S4, improves generalization across sequence modeling tasks with long-term dependencies such as image, text, audio, and medical time-series, with an average performance of 87.32\% on the Long-Range Arena benchmark. On the full raw Speech Command recognition dataset, Liquid-S4 achieves 96.78\% accuracy with a 30\% reduction in parameter counts compared to S4. The additional gain in performance is the direct result of the Liquid-S4's kernel structure that takes into account the similarities of the input sequence samples during training and inference.

****

Equivariant Hypergraph Diffusion Neural Operators

Peihao Wang,Shenghao Yang,Yunyu Liu,Zhangyang Wang,Pan Li

Hypergraph neural networks (HNNs) using neural networks to encode hypergraphs provide a promising way to model higher-order relations in data and further solve relevant prediction tasks built upon such higher-order relations. However, higher-order relations in practice contain complex patterns and are often highly irregular. So, it is often challenging to design an HNN that suffices to express those relations while keeping computational efficiency. Inspired by hypergraph diffusion algorithms, this work proposes a new HNN architecture named ED-HNN, which provably approximates any continuous equivariant hypergraph diffusion operators that can model a wide range of higher-order relations. ED-HNN can be implemented efficiently by combining star expansions of hypergraphs with standard message passing neural networks. ED-HNN further shows great superiority in processing heterophilic hypergraphs and constructing deep models. We evaluate ED-HNN for node classification on nine real-world hypergraph datasets. ED-HNN uniformly outperforms the best baselines over these nine datasets and achieves more than 2%$\uparrow$ in prediction accuracy over four datasets therein. Our code is available at: https://github.com/Graph-COM/ED-HNN.

****

Ollivier-Ricci Curvature for Hypergraphs: A Unified Framework

Corinna Coupette,Sebastian Dalleiger,Bastian Rieck

Bridging geometry and topology, curvature is a powerful and expressive invariant. While the utility of curvature has been theoretically and empirically confirmed in the context of manifolds and graphs, its generalization to the emerging domain of hypergraphs has remained largely unexplored. On graphs, the Ollivier-Ricci curvature measures differences between random walks via Wasserstein distances, thus grounding a geometric concept in ideas from probability theory and optimal transport. We develop Orchid, a flexible framework generalizing Ollivier-Ricci curvature to hypergraphs, and prove that the resulting curvatures have favorable theoretical properties. Through extensive experiments on synthetic and real-world hypergraphs from different domains, we demonstrate that Orchid curvatures are both scalable and useful to perform a variety of hypergraph tasks in practice.

****

Biases in Evaluation of Molecular Optimization Methods and Bias Reduction Strategies

Hiroshi Kajino,Kohei Miyaguchi,Takayuki Osogami

We are interested in in silico evaluation methodology for molecular optimization methods. Given a sample of molecules and their properties of our interest, we wish not only to train a generator of molecules that can find those optimized with respect to a target property but also to evaluate its performance accurately. A common practice is to train a predictor of the target property on the sample and use it for both training and evaluating the generator. We theoretically investigate this evaluation methodology and show that it potentially suffers from two biases; one is due to misspecification of the predictor and the other to reusing the same sample for training and evaluation. We discuss bias reduction methods for each of the biases, and empirically investigate their effectiveness.

****

Sharper Analysis of Sparsely Activated Wide Neural Networks with Trainable Biases

Hongru Yang,Ziyu Jiang,Ruizhe Zhang,Zhangyang Wang,Yingbin Liang

This work studies training one-hidden-layer overparameterized ReLU networks via gradient descent in the neural tangent kernel (NTK) regime, where, differently from the previous works, the networks' biases are trainable and are initialized to some constant rather than zero. The tantalizing benefit of such initialization is that the neural network will provably have sparse activation pattern before, during and after training, which can enable fast training procedures and, therefore, reduce the training cost. The first set of results of this work characterize the convergence of the network's gradient descent dynamics. The required width is provided to ensure gradient descent can drive the training error towards zero in linear rate. The contribution over previous work is that not only the bias is allowed to be updated by gradient descent under our setting but also a finer analysis is given such that the required width to ensure the network's closeness to its NTK is improved. Secondly, the networks' generalization bound after training is provided. A width-sparsity dependence is presented which yields sparsity-dependent localized Rademacher complexity and same-as-previous (ignoring logarithmic factors) generalization bound. Up to our knowledge, this is the first sparsity-dependent generalization result via localized Rademacher complexity. As a by-product, if the bias initialization is chosen to be zero, the width requirement improves the previous bound for the shallow networks' generalization. Lastly, since the generalization bound has dependence on the smallest eigenvalue of the limiting NTK and the bounds from previous works yield vacuous generalization, this work further studies the least eigenvalue of the limiting NTK. Surprisingly, while it is not shown that trainable biases are necessary, trainable bias helps to identify a nice data-dependent region where a much finer analysis of the NTK's smallest eigenvalue can be conducted, which leads to a much sharper lower bound than the previously known worst-case bound and, consequently, a non-vacuous generalization bound. Experimental evaluation is provided to evaluate our results.

****

Hard-Meta-Dataset++: Towards Understanding Few-Shot Performance on Difficult Tasks

Samyadeep Basu,Megan Stanley,John F Bronskill,Soheil Feizi,Daniela Massiceti

Few-shot classification is the ability to adapt to any new classification task from only a few training examples. The performance of current top-performing few-shot classifiers varies widely across different tasks where they often fail on a subset of `difficult' tasks.

This phenomenon has real-world consequences for deployed few-shot systems where safety and reliability are paramount, yet little has been done to understand these failure cases. In this paper, we study these difficult tasks to gain a more nuanced understanding of the limitations of current methods. To this end, we develop a general and computationally efficient algorithm called FastDiffSel to extract difficult tasks from any large-scale vision dataset. Notably, our algorithm can extract tasks at least 20x faster than existing methods enabling its use on large-scale datasets. We use FastDiffSel to extract difficult tasks from Meta-Da tasset, a widely-used few-shot classification benchmark, and other challenging large-scale vision datasets including ORBIT, CURE-OR and ObjectNet. These tasks are curated into Hard-MD++, a new few-shot testing benchmark to promote the development of methods that are robust to even the most difficult tasks. We use Hard-MD++ to stress-test an extensive suite of few-shot classification methods and show that state-of-the-art approaches fail catastrophically on difficult tasks. We believe that our extraction algorithm FastDiffSel and Hard-MD++ will aid researchers in further understanding failure modes of few-shot classification models.

****

REVISITING PRUNING AT INITIALIZATION THROUGH THE LENS OF RAMANUJAN GRAPH

Duc N.M Hoang,Shiwei Liu,Radu Marculescu,Zhangyang Wang

Pruning neural networks at initialization (PaI) has received an upsurge of interest due to its end-to-end saving potential. PaI is able to find sparse subnetworks at initialization that can achieve comparable performance to the full networks. These methods can surpass the trivial baseline of random pruning but suffer from a significant performance gap compared to post-training pruning. Previous approaches firmly rely on weights, gradients, and sanity checks as primary signals when conducting PaI analysis. To better understand the underlying mechanism of PaI, we propose to interpret it through the lens of the Ramanujan Graph - a class of expander graphs that are sparse while being highly connected. It is often believed there should be a strong correlation between the Ramanujan graph and PaI since both are about finding sparse and well-connected neural networks. However, the finer-grained link relating highly sparse and connected networks to their relative performance (i.e., ranking of difference sparse structures at the same specific global sparsity) is still missing. We observe that not only the Ramanujan property for sparse networks shows no significant relationship to PaI’s relative performance, but maximizing it can also lead to the formation of pseudo-random graphs with no structural meanings. We reveal the underlying cause to be Ramanujan Graph’s strong assumption on the upper bound of the largest nontrivial eigenvalue (μˆ) of layers belonging to highly sparse networks. We hence propose Iterative Mean Difference of Bound (IMDB) as a mean to relax the μˆ upper bound. Likewise, we also show there exists a lower bound for μˆ, which we call the Normalized Random Coefficient (NaRC), that gives us an accurate assessment for when sparse but highly connected structure degenerates into naive randomness. Finally, we systematically analyze the behavior of various PaI methods and demonstrate the utility of our proposed metrics in characterizing PaI performance. We show that subnetworks preserving better the IMDB property correlate higher in performance, while NaRC provides us with a possible mean to locate the region where highly connected, highly sparse, and non-trivial Ramanujan expanders exist. Our code is available at: https://github.com/VITA-Group/ramanujan-on-pai.

****

Self-supervised Speech Enhancement using Multi-Modal Data

Yu-Lin Wei,Bashima Islam,RAJALAXMI RAJAGOPALAN,romit choudhury

Modern earphones come equipped with microphones and inertial measurement units (IMU). When a user wears the earphone, the IMU can serve as a second modality for detecting speech signals. Specifically, as humans speak to their earphones (e.g., during phone calls), the throat’s vibrations propagate through the skull to ultimately induce a vibration in the IMU. The IMU data is heavily distorted (compared to the microphone’s recordings), but IMUs offer a critical advantage — they are not interfered by ambient sounds. This presents an opportunity in multi-modal speech enhancement, i.e., can the distorted but uninterfered IMU signal enhance the user’s speech when the microphone’s signal suffers from strong ambient interference?

We combine the best of both modalities (microphone and IMU) by designing a cooperative and self-supervised network architecture that does not rely on clean speech data from the user. Instead, using only noisy speech recordings, the IMU learns to give hints on where the target speech is likely located. The microphone uses this hint to enrich the speech signal, which then trains the IMU to improve subsequent hints. This iterative approach yields promising results, comparable to a supervised denoiser trained on clean speech signals. When clean signals are a lso available to our architecture, we observe promising SI-SNR improvement. We believe this result can aid speech-related applications in earphones and hearing aids, and potentially generalize to others, like audio-visual denoising.

****

Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers

Tianlong Chen,Zhenyu Zhang,AJAY KUMAR JAISWAL,Shiwei Liu,Zhangyang Wang

Despite their remarkable achievement, gigantic transformers encounter significant drawbacks, including exorbitant computational and memory footprints during training, as well as severe collapse evidenced by a high degree of parameter redundancy. Sparsely-activated Mixture-of-Experts (SMoEs) have shown promise to mitigate the issue of training efficiency, yet they are prone to (1) $\textit{redundant experts}$ due to representational collapse; and (2) $\textit{poor expert scalability for inference and downstream fine-tuning}$, primarily due to overfitting of the learned routing policy to the number of activated experts during training. As recent research efforts are predominantly focused on improving routing policies to encourage expert specializations, this work focuses on $\textit{exploring the overlooked scalability bottleneck of SMoEs}$ and leveraging it to effectively $\textbf{scale dense transformers}$. To this end, we propose a new plug-and-play training framework, $\textbf{SMoE-Dropout}$, to enable scaling transformers to better accuracy in their full capacity without collapse. Specifically, SMoE-Dropout consists of a $\textit{randomly initialized and fixed}$ router network to activate experts and gradually increases the activated expert number as training progresses over time. Transformers trained by SMoE-Dropout naturally exhibit a $\textbf{``self-slimmable”}$ property subject to resource availability, offering smooth and consistent performance boosts with an increase in activated experts during inference or fine-tuning. Our extensive experiments across diverse transformer architectures on a variety of tasks demonstrate the superior performance and substantial computation savings of SMoE-Dropout, compared to dense training baselines with equivalent parameter counts. In particular, our trained BERT outperforms its densely trained counterpart with consistent improvements of {1.03%, 0.78%, 1.09%} on challenging reasoning tasks {$\texttt{ASDiv-A}$, $\texttt{MAWPS}$, $\texttt{SVAMP}$}, respectively. Codes and models are available in https://github.com/VITA-Group/Random-MoE-as-Dropout.

****

Compositional Semantic Parsing with Large Language Models

Andrew Drozdov,Nathanael Schärli,Ekin Akyürek,Nathan Scales,Xinying Song,Xinyun Chen,Olivier Bousquet,Denny Zhou

Humans can reason compositionally when presented with new tasks. Previous research shows that appropriate prompting techniques enable large language models (LLMs) to solve artificial compositional generalization tasks such as SCAN. In this work, we identify additional challenges in more realistic semantic parsing tasks with larger vocabulary and refine these prompting techniques to address them. Our best method is based on least-to-most prompting: it decomposes the problem using prompting-based syntactic parsing, then uses this decomposition to select appropriate exemplars and to sequentially generate the semantic parse. This method allows us to set a new state of the art for CFQ while requiring only 1% of the training data used by traditional approaches. Due to the general nature of our approach, we expect similar efforts will lead to new results in other tasks and domains, especially for knowledge-intensive applications.

****

TiAda: A Time-scale Adaptive Algorithm for Nonconvex Minimax Optimization

Xiang Li,Junchi YANG,Niao He

Adaptive gradient methods have shown their ability to adjust the stepsizes on the fly in a parameter-agnostic manner, and empirically achieve faster convergence for solving minimization problems. When it comes to nonconvex minimax optimization, however, current convergence analyses of gradient descent ascent (GDA) combined with adaptive stepsizes require careful tuning of hyper-parameters and the knowledge of problem-dependent parameters. Such a discrepancy arises from the primal-dual nature of minimax problems and the necessity of delicate time-scale separation between the primal and dual updates in attaining convergence. In this w ork, we propose a single-loop adaptive GDA algorithm called TiAda for nonconvex minimax optimization that automatically adapts to the time-scale separation. Our algorithm is fully parameter-agnostic and can achieve near-optimal complexities simultaneously in deterministic and stochastic settings of nonconvex-strongly-concave minimax problems. The effectiveness of the proposed method is further justified numerically for a number of machine learning applications.

****

Generalization Properties of Retrieval-based Models

Soumya Basu,Ankit Singh Rawat,Manzil Zaheer

Many modern high-performing machine learning models such as GPT-3 primarily rely on scaling up models, e.g., transformer networks. Simultaneously, a parallel line of work aims to improve the model performance by augmenting an input instance with other (labeled) instances during inference. Examples of such augmentations include task-specific prompts and similar examples retrieved from the training data by a nonparametric component. Remarkably, retrieval-based methods have enjoyed success on a wide range of problems, ranging from standard natural language processing and vision tasks to protein folding, as demonstrated by many recent efforts, including WebGPT and AlphaFold. Despite growing literature showcasing the promise of these models, the theoretical underpinning for such models remains underexplored. In this paper, we present a formal treatment of retrieval-based models to characterize their generalization ability. In particular, we focus on two classes of retrieval-based classification approaches: First, we analyze a local learning framework that employs an explicit local empirical risk minimization based on retrieved examples for each input instance. Interestingly, we show that breaking down the underlying learning task into local sub-tasks enables the model to employ a low complexity parametric component to ensure good overall accuracy. The second class of retrieval-based approaches we explore learns a global model using kernel methods to directly map an input instance and retrieved examples to a prediction, without explicitly solving a local learning task.

****

Semi-Variance Reduction for Fair Federated Learning

Saber Malekmohammadi,Guojun Zhang,Yaoliang Yu

Ensuring fairness in Federated Learning (FL) systems, i.e. ensuring a satisfactory performance for all of the diverse clients in the systems, is an important and challenging problem. There are multiple fair FL algorithms in the literature, which have been relatively successful in providing fairness. However, these algorithms mostly emphasize on the loss functions of worst-off clients to improve their performance, which often results in the suppression of well-performing ones. As a consequence, the system's overall average performance is usually sacrificed for achieving fairness. Motivated by this and inspired by two well-known risk modeling methods in Finance, Mean-Variance and Mean-Semi-Variance, we propose and study two new fair FL algorithms, Variance Reduction (VRed) and Semi-Variance Reduction (Semi-VRed). VRed encourages equality between clients loss functions by penalizing their variance. In contrast, Semi-VRed penalizes the discrepancy of only the worst-off clients loss functions from the average loss. Through extensive experiments on multiple vision and language datasets, we show that, Semi-VRed achieves SoTA performance in scenarios with highly heterogeneous data distributions by improving both fairness and the system overall average performance at the same time.

****

Multi-Modality Alone is Not Enough: Generating Scene Graphs using Cross-Relation-Modality Tokens

Gopika Sudhakaran,Devendra Singh Dhami,Stefan Roth,Kristian Kersting

Recent years have seen a growing interest in Scene Graph Generation (SGG), a comprehensive visual scene understanding task that aims to predict the relationships between objects detected in a scene. One of its key challenges is the strong bias of the visual world around us toward a few frequently occurring relationships, leaving a long tail of under-represented classes. Although infusing additional modalities is one prominent way to improve SGG performance on under-represented classes, we argue that using additional modalities alone is not enough. We pro pose to inject entity relation information (Cross-Relation) and modality dependencies (Cross-Modality) into each embedding token of a transformer which we term primal fusion. The resulting Cross-RElAtion-Modality (CREAM) token acts as a strong inductive bias for the SGG framework. Our experimental results on the Visual Genome dataset demonstrate that our CREAM model outperforms state-of-the-art SGG models by around 20% while being simpler and requiring substantially less computation. Additionally, to analyse the generalisability of the CREAM model we also evaluate it on the Open Images dataset. Finally, we examine the impact of the depth-map quality on SGG performance and empirically show the superiority of our model over the prior state of the art by better capturing the depth data, boosting the performance by a margin of around 25%.

****

FaiREE: fair classification with finite-sample and distribution-free guarantee
Puheng Li,James Zou,Linjun Zhang

Algorithmic fairness plays an increasingly critical role in machine learning research. Several group fairness notions and algorithms have been proposed. However, the fairness guarantee of existing fair classification methods mainly depend on specific data distributional assumptions, often requiring large sample sizes, and fairness could be violated when there is a modest number of samples, which is often the case in practice. In this paper, we propose FaiREE, a fair classification algorithm which can satisfy group fairness constraints with finite-sample and distribution-free theoretical guarantees. FaiREE can be adapted to satisfying various group fairness notions (e.g., Equality of Opportunity, Equalized Odds, Demographic Parity, etc.) and achieve the optimal accuracy. These theoretical guarantees are further supported by experiments on both synthetic and real data. FaiREE is shown to have favorable performance over state-of-the-art algorithms.

****

Deep Evidential Reinforcement Learning for Dynamic Recommendations
Dingrong Wang,Krishna Prasad Neupane,Ervine Zheng,Qi Yu

Reinforcement learning (RL) has been applied to build recommender systems (RS) to capture users' evolving preferences and continuously improve the quality of recommendations. In this paper, we propose a novel deep evidential reinforcement learning (DERL) framework that learns a more effective recommendation policy by integrating both the expected reward and evidence-based uncertainty. In particular, DERL conducts evidence-aware exploration to locate items that a user will most likely take interest in the future. Two central components of DERL include a customized recurrent neural network (RNN) and an evidential-actor-critic (EAC) module. The former module is responsible for generating the current state of the environment by aggregating historical information and a sliding window that contains the current user interactions as well as newly recommended items that may encode future interest. The latter module performs evidence-based exploration by maximizing a uniquely designed evidential Q-value to derive a policy giving preference to items with good predicted ratings while remaining largely unknown to the system (due to lack of evidence). These two components are jointly trained by supervised learning and reinforcement learning. Experiments on multiple real-world dynamic datasets demonstrate the state-of-the-art performance of DERL and its capability to capture long-term user interests.

****

Exponential Generalization Bounds with Near-Optimal Rates for L_q-Stable Algorithms
Xiaotong Yuan,Ping Li

The \emph{stability} of learning algorithms to changes in the training sample has been actively studied as a powerful proxy for reasoning about generalization. Recently, exponential generalization and excess risk bounds with near-optimal rates have been obtained under the stringent and distribution-free notion of uniform stability~\citep{bousquet2020sharper,klochkov2021stability}. In the meanwhile, under the notion of L_q-stability, which is weaker and distribution dependent, exponential generalization bounds are also available yet so far only with sub-optimal rates. Therefore, a fundamental question we would like to address in this paper is whether it is possible to derive near-optimal exponential generaliz

ation bounds for L_q-stable learning algorithms. As the core contribution of the present work, we give an affirmative answer to this question by developing strict analogues of the near-optimal generalization and risk bounds of uniformly stable algorithms for L_q-stable algorithms. Further, we demonstrate the power of our improved L_q-stability and generalization theory by applying it to derive strong sparse excess risk bounds, under mild conditions, for computationally tractable sparsity estimation algorithms such as Iterative Hard Thresholding (IHT).
****

Disentangling Learning Representations with Density Estimation
Eric Yeats,Frank Y Liu,Hai Li
Disentangled learning representations have promising utility in many applications, but they currently suffer from serious reliability issues. We present Gaussian Channel Autoencoder (GCAE), a method which achieves reliable disentanglement via scalable non-parametric density estimation of the latent space. GCAE avoids the curse of dimensionality of density estimation by disentangling subsets of its latent space with the Dual Total Correlation (DTC) metric, thereby representing its high-dimensional latent joint distribution as a collection of many low-dimensional conditional distributions. In our experiments, GCAE achieves highly competitive and reliable disentanglement scores compared with state-of-the-art baselines.
****

Coarse-to-fine Knowledge Graph Domain Adaptation based on Distantly-supervised Iterative Training
Hongmin Cai,Wenxiong Liao,Zhengliang Liu,Yuzhong Chen,Tianming Liu,Xiang Li
Modern supervised learning neural network models require a large amount of manually labeled data, which makes the construction of domain-specific knowledge graphs time-consuming and labor-intensive. In parallel, although there has been much research on named entity recognition and relation extraction based on distantly supervised learning, constructing a domain-specific knowledge graph from large collections of textual data without manual annotations is still an urgent problem to be solved. In response, we propose an integrated framework for adapting and re-learning knowledge graphs from one coarse domain (biomedical) to a finer-define domain (oncology). In this framework, we apply distant-supervision on cross-domain knowledge graph adaptation. Consequently, no manual data annotation is required to train the model. We introduce a novel iterative training strategy to facilitate the discovery of domain-specific named entities and triples. Experimental results indicate that the proposed framework can perform domain adaptation and construction of knowledge graph efficiently.
****

Teacher Guided Training: An Efficient Framework for Knowledge Transfer
Manzil Zaheer,Ankit Singh Rawat,Seungyeon Kim,Chong You,Himanshu Jain,Andreas Veit,Rob Fergus,Sanjiv Kumar
The remarkable performance gains realized by large pretrained models, e.g., GPT-3, hinge on the massive amounts of data they are exposed to during training. Analogously, distilling such large models to compact models for efficient deployment also necessitates a large amount of (labeled or unlabeled) training data. In this paper, we propose the teacher-guided training (TGT) framework for training a high-quality compact model that leverages the knowledge acquired by pretrained generative models, while obviating the need to go through a large volume of data. TGT exploits the fact that the teacher has acquired a good representation of the underlying data domain, which typically corresponds to a much lower dimensional manifold than the input space. Furthermore, we can use the teacher to explore input space more efficiently through sampling or gradient-based methods; thus, making TGT especially attractive for limited data or long-tail settings. We formally capture this benefit of proposed data-domain exploration in our generalization bounds. We find that TGT can improve accuracy on several image classification benchmarks as well as a range of text classification and retrieval tasks.
****

Neural Agents Struggle to Take Turns in Bidirectional Emergent Communication

Valentin Taillandier,Dieuwke Hupkes,Benoît Sagot,Emmanuel Dupoux,Paul Michel

The spontaneous exchange of turns is a central aspect of human communication. Although turn-taking conventions come to us naturally, artificial dialogue agents struggle to coordinate, and must rely on hard-coded rules to engage in interactive conversations with human interlocutors. In this paper, we investigate the conditions under which artificial agents may naturally develop turn-taking conventions in a simple language game. We describe a cooperative task where success is contingent on the exchange of information along a shared communication channel where talking over each other hinders communication. Despite these environmental constraints, neural-network based agents trained to solve this task with reinforcement learning do not systematically adopt turn-taking conventions. However, we find that agents that do agree on turn-taking protocols end up performing better.

Moreover, agents that are forced to perform turn-taking can learn to solve the task more quickly.

This suggests that turn-taking may help to generate conversations that are easier for speakers to interpret.

****

Class Interference of Deep Networks

Dongcui Diao,Hengshuai Yao,Bei Jiang

Recognizing and telling similar objects apart is even hard for human beings. In this paper, we show that there is a phenomenon of class interference with all deep neural networks. Class interference represents the learning difficulty in data and it constitutes the largest percentage of generalization errors by deep networks. To understand class interference, we propose cross-class tests, class ego directions and interference models. We show how to use these definitions to study minima flatness and class interference of a trained model. We also show how to detect class interference during training through label dancing pattern and class dancing notes.

****

Observational Robustness and Invariances in Reinforcement Learning via Lexicographic Objectives

Daniel Jarne Ornia,Licio Romao,Lewis Hammond,Manuel Mazo Jr,Alessandro Abate

Policy robustness in Reinforcement Learning (RL) may not be desirable at any price; the alterations caused by robustness requirements from otherwise optimal policies should be explainable and quantifiable. Policy gradient algorithms that have strong convergence guarantees are usually modified to obtain robust policies in ways that do not preserve algorithm guarantees, which defeats the purpose of formal robustness requirements. In this work we study a notion of robustness in partially observable MDPs where state observations are perturbed by a noise-induced stochastic kernel. We characterize the set of policies that are maximally robust by analysing how the policies are altered by this kernel. We then establish a connection between such robust policies and certain properties of the noise kernel, as well as with structural properties of the underlying MDPs, constructing sufficient conditions for policy robustness. We use these notions to propose a robustness-inducing scheme, applicable to any policy gradient algorithm, to formally trade off the reward achieved by a policy with its robustness level through lexicographic optimization, which preserves convergence properties of the original algorithm. We test the the proposed approach through numerical experiments on safety-critical RL environments, and show how the proposed method helps achieve high robustness when state errors are introduced in the policy roll-out.

****

SeedGNN: Graph Neural Network for Supervised Seeded Graph Matching

Liren Yu,Jiaming Xu,Xiaojun Lin

There have been significant interests in designing Graph Neural Networks (GNNs) for seeded graph matching, which aims to match two (unlabeled) graphs using only topological information and a small set of seeds. However, most previous GNNs for seeded graph matching employ a semi-supervised approach, which requires a large number of seeds and can not learn knowledge transferable to unseen graphs. In contrast, this paper proposes a new supervised approach that can learn from a t

raining set how to match unseen graphs with only a few seeds. At the core of our SeedGNN architecture are two novel modules: 1) a convolution module that can easily learn the capability of counting and using witnesses of different hops; 2) a percolation module that can use easily-matched pairs as new seeds to percolate and match other nodes. We evaluate SeedGNN on both synthetic and real graphs, and demonstrate significant performance improvement over both non-learning and learning algorithms in the existing literature. Further, our experiments confirm that the knowledge learned by SeedGNN from training graphs can be generalized to test graphs with different sizes and categories.
****

Provable Sharpness-Aware Minimization with Adaptive Learning Rate
Hao Sun,Li Shen,Qihuang Zhong,Liang Ding,Shixiang Chen,Jingwei Sun,Guangzhong Sun,Dacheng Tao
Sharpness aware minimization (SAM) optimizer has been extensively explored as it can converge fast and train deep neural networks efficiently via introducing extra perturbation steps to flatten the landscape of deep learning models. A combination of SAM with adaptive learning rate (AdaSAM) has also been explored to train large-scale deep neural networks without theoretical guarantee due to the dual difficulties in analyzing the perturbation step and the coupled adaptive learning rate. In this paper, we try to analyze the convergence rate of AdaSAM in the stochastic non-convex setting. We theoretically show that AdaSAM admit a $\mathcal{O}(1/\sqrt{bT})$ convergence rate and show linear speedup property with respect to mini-batch size b. To best of our knowledge, we are the first to provide the non-trivial convergence rate of SAM with an adaptive learning rate. To decouple the two stochastic gradient steps with the adaptive learning rate, we first introduce the delayed second-order momentum during the convergence to decompose them to make them independent while taking an expectation. Then we bound them by showing the adaptive learning rate has a limited range, which makes our analysis feasible. At last, we conduct experiments on several NLP tasks and they show that AdaSAM could achieve superior performance compared with SGD, AMSGrad, and SAM optimizer.
****

Prompting GPT-3 To Be Reliable
Chenglei Si,Zhe Gan,Zhengyuan Yang,Shuohang Wang,Jianfeng Wang,Jordan Lee Boyd-Graber,Lijuan Wang
Large language models (LLMs) show impressive abilities via few-shot prompting. Commercialized APIs such as OpenAI GPT-3 further increase their use in real-world language applications. However, the crucial problem of how to improve the reliability of GPT-3 is still under-explored. While reliability is a broad and vaguely defined term, we decompose reliability into four main facets that correspond to the existing framework of ML safety and are well-recognized to be important: generalizability, social biases, calibration, and factuality. Our core contribution is to establish simple and effective prompts that improve GPT-3’s reliability as it: 1) generalizes out-of-distribution, 2) balances demographic distribution and uses natural language instructions to reduce social biases, 3) calibrates output probabilities, and 4) updates the LLM’s factual knowledge and reasoning chains. With appropriate prompts, GPT-3 is more reliable than smaller-scale supervised models on all these facets. We release all processed datasets, evaluation scripts, and model predictions. Our systematic empirical study not only sheds new insights on the reliability of prompting LLMs, but more importantly, our prompting strategies can help practitioners more reliably use LLMs like GPT-3.
****

Adversarial Training of Self-supervised Monocular Depth Estimation against Physical-World Attacks
Zhiyuan Cheng,James Chenhao Liang,Guanhong Tao,Dongfang Liu,Xiangyu Zhang
Monocular Depth Estimation (MDE) is a critical component in applications such as autonomous driving. There are various attacks against MDE networks. These attacks, especially the physical ones, pose a great threat to the security of such systems. Traditional adversarial training method requires ground-truth labels and hence cannot be directly applied to self-supervised MDE that does not have dept

h ground truth. Some self-supervised model hardening technique (e.g., contrastive learning) ignores the domain knowledge of MDE and can hardly achieve optimal performance. In this work, we propose a novel adversarial training method for self-supervised MDE models based on view synthesis without using the depth ground truth. We improve adversarial robustness against physical-world attacks using L_0-norm-bounded perturbation in training. We compare our method with supervised learning-based and contrastive learning-based methods that are tailored for MDE. Results on two representative MDE networks show that we achieve better robustness against various adversarial attacks with nearly no benign performance degradation.

****

Sparsity-Constrained Optimal Transport

Tianlin Liu,Joan Puigcerver,Mathieu Blondel

Regularized optimal transport (OT) is now increasingly used as a loss or as a matching layer in neural networks. Entropy-regularized OT can be computed using the Sinkhorn algorithm but it leads to fully-dense transportation plans, meaning that all sources are (fractionally) matched with all targets. To address this issue, several works have investigated quadratic regularization instead. This regularization preserves sparsity and leads to unconstrained and smooth (semi) dual objectives, that can be solved with off-the-shelf gradient methods. Unfortunately, quadratic regularization does not give direct control over the cardinality (number of nonzeros) of the transportation plan. We propose in this paper a new approach for OT with explicit cardinality constraints on the transportation plan. Our work is motivated by an application to sparse mixture of experts, where OT can be used to match input tokens such as image patches with expert models such as neural networks. Cardinality constraints ensure that at most k tokens are matched with an expert, which is crucial for computational performance reasons. Despite the nonconvexity of cardinality constraints, we show that the corresponding (semi) dual problems are tractable and can be solved with first-order gradient methods. Our method can be thought as a middle ground between unregularized OT (recovered in the limit case $k=1$) and quadratically-regularized OT (recovered when k is large enough). The smoothness of the objectives increases as k increases, giving rise to a trade-off between convergence speed and sparsity of the optimal plan.

****

A Risk-Averse Equilibrium for Multi-Agent Systems

Oliver Slumbers,David Henry Mguni,Stephen Marcus McAleer,Jun Wang,Yaodong Yang

In multi-agent systems, intelligent agents are tasked with making decisions that lead to optimal outcomes when actions of the other agents are as expected, whilst also being prepared for their unexpected behaviour. In this work, we introduce a novel risk-averse solution concept that allows the learner to accommodate low probability actions by finding the strategy with minimum variance, given any level of expected utility. We first prove the existence of such a risk-averse equilibrium, and propose one fictitious-play type learning algorithm for smaller games that enjoys provable convergence guarantees in games classes including zero-sum and potential. Furthermore, we propose an approximation method for larger games based on iterative population-based training that generates a population of risk- averse agents. Empirically, our equilibrium is shown to be able to reduce the utility variance, specifically in the sense that other agents’ low probability behaviour is better accounted for by our equilibrium in comparison to playing other solutions. Importantly, we show that our population of agents that approximate a risk-averse equilibrium is particularly effective against unseen opposing populations, especially in the case of guaranteeing a minimum level of performance, which is critical to safety-aware multi-agent systems.

****

SuperWeight Ensembles: Automated Compositional Parameter Sharing Across Diverse Architechtures

Piotr Teterwak,Soren Nelson,Nikoli Dryden,Dina Bashkirova,Kate Saenko,Bryan A. Plummer

Neural net ensembles boost task performance, but have excessive storage requirem ents. Recent work in efficient ensembling has made the memory cost more tractable by sharing learned parameters between ensemble members. Existing efficient ensembles have high predictive accuracy, but they are overly restrictive in two ways: 1) They constrain ensemble members to have the same architecture, limiting their usefulness in applications such as anytime inference, and 2) They reduce the parameter count for a small predictive performance penalty, but do not provide an easy way to trade-off parameter count for predictive performance without increasing inference time. In this paper, we propose SuperWeight Ensembles, an approach for architecture-agnostic parameter sharing. SuperWeight Ensembles share parameters between layers which have sufficiently similar computation, even if they have different shapes. This allows anytime prediction of heterogeneous ensembles by selecting a subset of members during inference, which is a flexibility not supported by prior work. In addition, SuperWeight Ensembles provide control over the total number of parameters used, allowing us to increase or decrease the number of parameters without changing model architecture. On the anytime prediction task, our method shows a consistent boost over prior work while allowing for more flexibility in architectures and efficient parameter sharing. SuperWeight Ensembles preserve the performance of prior work in the low-parameter regime, and even outperform fully-parameterized ensembles with 17% fewer parameters on CIFAR-100 and 50% fewer parameters on ImageNet.

****

Human alignment of neural network representations

Lukas Muttenthaler,Jonas Dippel,Lorenz Linhardt,Robert A. Vandermeulen,Simon Kornblith

Today’s computer vision models achieve human or near-human level performance across a wide variety of vision tasks. However, their architectures, data, and learning algorithms differ in numerous ways from those that give rise to human vision. In this paper, we investigate the factors that affect the alignment between the representations learned by neural networks and human mental representations inferred from behavioral responses. We find that model scale and architecture have essentially no effect on the alignment with human behavioral responses, whereas the training dataset and objective function both have a much larger impact. These findings are consistent across three datasets of human similarity judgments collected using two different tasks. Linear transformations of neural network representations learned from behavioral responses from one dataset substantially improve alignment with human similarity judgments on the other two datasets. In addition, we find that some human concepts such as food and animals are well-represented by neural networks whereas others such as royal or sports-related objects are not. Overall, although models trained on larger, more diverse datasets achieve better alignment with humans than models trained on ImageNet alone, our results indicate that scaling alone is unlikely to be sufficient to train neural networks with conceptual representations that match those used by humans.

****

Imitation Learning for Mean Field Games with Correlated Equilibria

Zhiyu Zhao,Renyuan Xu,Haifeng Zhang,Jun Wang,Mingyuan Zhang,Yaodong Yang

Imitation learning (IL) aims at achieving optimal actions by learning from demonstrated behaviors without knowing the reward function and transition kernels. Conducting IL with a large population of agents is challenging as agents' interactions grow exponentially with respect to the population size. Mean field theory provides an efficient tool to study multi-agent problems by aggregating information on the population level. While the approximation is tractable, it is non-trivial to restore mean field Nash equilibria (MFNE) from demonstrations. Importantly, there are many real-world problems that cannot be explained by the classic MFNE concept; this includes the traffic network equilibrium induced from the public routing recommendations and the pricing equilibrium of goods generated on the E-commerce platform.   In both examples, correlated devices are introduced to the equilibrium due to the intervention from the platform. To accommodate this, we propose a novel solution concept named adaptive mean field correlated equilibrium (AMFCE) that generalizes MFNE. On the theory side, we first prove the existence of AMFCE, and establish a novel framework based on IL and AMFCE with entropy regularization (MaxEnt-AMFCE) to recover the AMFCE policy from real-world demonstrations. Signatures from the rough path theory are then applied to characterize the mean-field evolution. A significant benefit of MaxEnt-AMFCE is that it can recover both the equilibrium policy and the correlation device from data. We test our MaxEnt-AMFCE against the state-of-the-art IL algorithms for MFGs on several tasks (including a real-world traffic flow prediction problem), results justify the effectiveness of our proposed method and show its potential to predicting and explaining large population behavior under correlated signals.

****

Turning the Curse of Heterogeneity in Federated Learning into a Blessing for Out-of-Distribution Detection

Shuyang Yu,Junyuan Hong,Haotao Wang,Zhangyang Wang,Jiayu Zhou

Deep neural networks have witnessed huge successes in many challenging prediction tasks and yet they often suffer from out-of-distribution (OoD) samples, misclassifying them with high confidence. Recent advances show promising OoD detection performance for centralized training, and however, OoD detection in federated learning (FL) is largely overlooked, even though many security sensitive applications such as autonomous driving and voice recognition authorization are commonly trained using FL for data privacy concerns. The main challenge that prevents previous state-of-the-art OoD detection methods from being incorporated to FL is that they require large amount of real OoD samples. However, in real-world scenarios, such large-scale OoD training data can be costly or even infeasible to obtain, especially for resource-limited local devices. On the other hand, a notorious challenge in FL is data heterogeneity where each client collects non-identically and independently distributed (non-iid) data. We propose to take advantage of such heterogeneity and turn the curse into a blessing that facilitates OoD detection in FL. The key is that for each client, non-iid data from other clients (unseen external classes) can serve as an alternative to real OoD samples. Specifically, we propose a novel Federated Out-of-Distribution Synthesizer (FOSTER), which learns a class-conditional generator to synthesize virtual external-class OoD samples, and maintains data confidentiality and communication efficiency required by FL. Experimental results show that our method outperforms the state-of-the-art by 2.49%, 2.88%, 1.42% AUROC, and 0.01%, 0.89%, 1.74% ID accuracy, on CIFAR-10, CIFAR-100, and STL10, respectively.

****

Clustering and Ordering Variable-Sized Sets: The Catalog Problem

Mateusz Maria Jurewicz,Graham W. Taylor,Leon Derczynski

Prediction of a varying number of ordered clusters from sets of any cardinality is a challenging task for neural networks, combining elements of set representation, clustering and learning to order. This task arises in many diverse areas, ranging from medical triage, through multi-channel signal analysis for petroleum exploration to product catalog structure prediction. This paper focuses on the latter, which exemplifies a number of challenges inherent to adaptive ordered clustering, referred to further as the eponymous Catalog Problem. These include learning variable cluster constraints, exhibiting relational reasoning and managing combinatorial complexity. Despite progress in both neural clustering and set-to-sequence methods, no joint, fully differentiable model exists to-date. We develop such a modular architecture, referred to further as Neural Ordered Clusters (NOC), enhance it with a specific mechanism for learning cluster-level cardinality constraints, and provide a robust comparison of its performance in relation to alternative models. We test our method on three datasets, including synthetic catalog structures and PROCAT, a dataset of real-world catalogs consisting of over 1.5M products, achieving state-of-the-art results on a new, more challenging formulation of the underlying problem, which has not been addressed before. Additionally, we examine the network's ability to learn higher-order interactions and investigate its capacity to learn both compositional and structural rulesets.

****

RangeAugment: Efficient Online Augmentation with Range Learning

Sachin Mehta,Saeid Naderiparizi,Fartash Faghri,Maxwell Horton,Lailin Chen,Ali Farhadi,Oncel Tuzel,Mohammad Rastegari

State-of-the-art automatic augmentation methods (e.g., AutoAugment and RandAugment) for visual recognition tasks diversify training data using a large set of augmentation operations. The range of magnitudes of many augmentation operations (e.g., brightness and contrast) is continuous. Therefore, to make search computationally tractable, these methods use fixed and manually-defined magnitude ranges for each operation, which may lead to sub-optimal policies. To answer the open question on the importance of magnitude ranges for each augmentation operation, we introduce RangeAugment that allows us to efficiently learn the range of magnitudes for individual as well as composite augmentation operations. RangeAugment uses an auxiliary loss based on image similarity as a measure to control the range of magnitudes of augmentation operations. As a result, RangeAugment has a single scalar parameter for search, image similarity, which we simply optimize via linear search. RangeAugment integrates seamlessly with any model and learns model- and task-specific augmentation policies. With extensive experiments on the ImageNet dataset across different networks, we show that RangeAugment achieves competitive performance to state-of-the-art automatic augmentation methods with 4-5 times fewer augmentation operations. Experimental results on semantic segmentation and contrastive learning further shows RangeAugment's effectiveness.

****

How Predictors Affect Search Strategies in Neural Architecture Search?

TianJin Deng,Jia Wu

Predictor-based Neural Architecture Search (NAS) is an important topic since it can efficiently reduce the computational cost of evaluating candidate architectures. Most existing predictor-based NAS algorithms aim to design different predictors to improve prediction performance. Unfortunately, even a promising performance predictor may suffer from the accuracy decline due to long-term and continuous usage, thus leading to the degraded performance of the search strategy. That naturally gives rise to the following problems: how predictors affect search strategies and how to appropriately use the predictor? In this paper, we take reinforcement learning (RL) based search strategy to study theoretically and empirically the impact of predictors on search strategies. We first formulate a predictor-RL-based NAS algorithm as model-based RL and analyze it with a guarantee of monotonic improvement at each trail. Then, based on this analysis, we propose a simple procedure of predictor usage, named mixed batch, which contains ground-truth data and prediction data. The proposed procedure can efficiently reduce the impact of predictor errors on search strategies with maintaining performance growth. Our algorithm, Predictor-based Neural Architecture Search with Mixed batch (PNASM), outperforms traditional NAS algorithms and prior state-of-the-art predictor-based NAS algorithms on three NAS-Bench-201 tasks.

****

Unbiased Stochastic Proximal Solver for Graph Neural Networks with Equilibrium States

Mingjie Li,Yifei Wang,Yisen Wang,Zhouchen Lin

Graph Neural Networks (GNNs) are widely used deep learning models that can extract meaningful representations from graph datasets and achieve great success in many machine learning tasks. Among them, graph neural networks with iterative iterations like unfolded GNNs and implicit GNNs can effectively capture long-range dependencies in graphs and demonstrate superior performance on large graphs since they can mathematically ensure its convergence to some nontrivial solution after lots of aggregations. However, the aggregation time for such models costs a lot as they need to aggregate the full graph in each update. Such weakness limits the scalability of the implicit graph models. To tackle such limitations, we propose two unbiased stochastic proximal solvers inspired by the stochastic proximal gradient descent method and its variance reduction variant called USP and USP-VR solvers. From the point of stochastic optimization, we theoretically prove that our solvers are unbiased, which can converge to the same solution as the original solvers for unfolded GNNs and implicit GNNs. Furthermore, the computation complexities for unfolded GNNs and implicit GNNs with our proposed solvers are significantly less than their vanilla versions. Experiments on various large grap

h datasets show that our proposed solvers are more efficient and can achieve state-of-the-art performance.
****

Energy Transformer

Benjamin Hoover,Yuchen Liang,Bao Pham,Rameswar Panda,Hendrik Strobelt,Duen Horng Chau,Mohammed J Zaki,Dmitry Krotov

Transformers have become the de facto models of choice in machine learning, typically leading to impressive performance on many applications. At the same time, the architectural development in the transformer world is mostly driven by empirical findings, and the theoretical understanding of their architectural building blocks is rather limited. In contrast, Dense Associative Memory models or Modern Hopfield Networks have a well-established theoretical foundation, but have not yet demonstrated truly impressive practical results. We propose a transformer architecture that replaces the sequence of feedforward transformer blocks with a single large Associative Memory model. Our novel architecture, called Energy Transformer (or ET for short), has many of the familiar architectural primitives that are often used in the current generation of transformers. However, it is not identical to the existing architectures. The sequence of transformer layers in ET is purposely designed to minimize a specifically engineered energy function, which is responsible for representing the relationships between the tokens. As a consequence of this computational principle, the attention in ET is different from the conventional attention mechanism. In this work, we introduce the theoretical foundations of ET, explore it's empirical capabilities using the image completion task, and obtain strong quantitative results on the graph anomaly detection task.
****

Privacy-Preserving Vision Transformer on Permutation-Encrypted Images

Fusheng Hao,Fengxiang He,Yikai Wang,Fuxiang Wu,Jun Cheng,Dacheng Tao

Massive human-related data is collected to train neural networks for computer vision tasks. Potential incidents, such as data leakages, expose significant privacy risks to applications. In this paper, we propose an efficient privacy-preserving learning paradigm, where images are first encrypted via one of the two encryption strategies: (1) random shuffling to a set of equally-sized patches and (2) mixing-up sub-patches of the images. Then, a permutation-equivariant vision transformer is designed to learn on the encrypted images for vision tasks, including image classification and object detection. Extensive experiments on ImageNet and COCO show that the proposed paradigm achieves comparable accuracy with the competitive methods. Moreover, decrypting the encrypted images is solving an NP-hard jigsaw puzzle or an ill-posed inverse problem, which is empirically shown intractable to be recovered by the powerful vision transformer-based attackers. We thus show that the proposed paradigm can destroy human-recognizable contents while preserving machine-learnable information. Code will be released publicly.
****

DiGress: Discrete Denoising diffusion for graph generation

Clement Vignac,Igor Krawczuk,Antoine Siraudin,Bohan Wang,Volkan Cevher,Pascal Frossard

This work introduces DiGress, a discrete denoising diffusion model for generating graphs with categorical node and edge attributes.

Our model utilizes a discrete diffusion process that progressively edits graphs with noise, through the process of adding or removing edges and changing the categories.

A graph transformer network is trained to revert this process, simplifying the problem of distribution learning over graphs into a sequence of node and edge classification tasks.

We further improve sample quality by introducing a Markovian noise model that preserves the marginal distribution of node and edge types during diffusion, and by incorporating auxiliary graph-theoretic features.

A procedure for conditioning the generation on graph-level features is also proposed.

DiGress achieves state-of-the-art performance on molecular and non-molecular dat

asets, with up to 3x validity improvement on a planar graph dataset.
It is also the first model to scale to the large GuacaMol dataset containing 1.3M drug-like molecules without the use of molecule-specific representations.

****

DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion

Qitian Wu,Chenxiao Yang,Wentao Zhao,Yixuan He,David Wipf,Junchi Yan

Real-world data generation often involves complex inter-dependencies among instances, violating the IID-data hypothesis of standard learning paradigms and posing a challenge for uncovering the geometric structures for learning desired instance representations. To this end, we introduce an energy constrained diffusion model which encodes a batch of instances from a dataset into evolutionary states that progressively incorporate other instances' information by their interactions. The diffusion process is constrained by descent criteria w.r.t.~a principled energy function that characterizes the global consistency of instance representations over latent structures. We provide rigorous theory that implies closed-form optimal estimates for the pairwise diffusion strength among arbitrary instance pairs, which gives rise to a new class of neural encoders, dubbed as DIFFormer, with two instantiations: a simple version with linear complexity for prohibitive instance numbers, and an advanced version for learning complex structures. Experiments highlight the wide applicability of our model as a general-purpose encoder backbone with superior performance in various tasks, such as semi-supervised node classification, image/text classification, and spatial-temporal dynamics prediction.

****

Neural Lagrangian Schr\"{o}dinger Bridge: Diffusion Modeling for Population Dynamics

Takeshi Koshizuka,Issei Sato

Population dynamics is the study of temporal and spatial variation in the size of populations of organisms and is a major part of population ecology. One of the main difficulties in analyzing population dynamics is that we can only obtain observation data with coarse time intervals from fixed-point observations due to experimental costs or measurement constraints. Recently, modeling population dynamics by using continuous normalizing flows (CNFs) and dynamic optimal transport has been proposed to infer the sample trajectories from a fixed-point observed population. While the sample behavior in CNFs is deterministic, the actual sample in biological systems moves in an essentially random yet directional manner. Moreover, when a sample moves from point A to point B in dynamical systems, its trajectory typically follows the principle of least action in which the corresponding action has the smallest possible value. To satisfy these requirements of the sample trajectories, we formulate the Lagrangian Schrödinger bridge (LSB) problem and propose to solve it approximately by modeling the advection-diffusion process with regularized neural SDE. We also develop a model architecture that enables faster computation of the loss function. Experimental results show that the proposed method can efficiently approximate the population-level dynamics even for high-dimensional data and that using the prior knowledge introduced by the Lagrangian enables us to estimate the sample-level dynamics with stochastic behavior.

****

Jump-Start Reinforcement Learning

Ikechukwu Uchendu,Ted Xiao,Yao Lu,Banghua Zhu,Mengyuan Yan,Joséphine Simon,Matthew Bennice,Chuyuan Fu,Cong Ma,Jiantao Jiao,Sergey Levine,Karol Hausman

Reinforcement learning (RL) provides a theoretical framework for continuously improving an agent’s behavior via trial and error. However, efficiently learning policies from scratch can be very difficult, particularly for tasks that present exploration challenges. In such settings, it might be desirable to initialize RL with an existing policy, offline data, or demonstrations. However, naively performing such initialization in RL often works poorly, especially for value-based methods. In this paper, we present a meta algorithm that can use offline data, demonstrations, or a pre-existing policy to initialize an RL policy, and is compatible with any RL approach. In particular, we propose Jump-Start Reinforcement L earning (JSRL), an algorithm that employs two policies to solve tasks: a guide-policy, and an exploration-policy. By using the guide-policy to form a curriculum of starting states for the exploration-policy, we are able to efficiently improve performance on a set of simulated robotic tasks. We show via experiments that it is able to significantly outperform existing imitation and reinforcement learning algorithms, particularly in the small-data regime. In addition, we provide an upper bound on the sample complexity of JSRL and show that with the help of a guide-policy, one can improve the sample complexity for non-optimism exploration methods from exponential in horizon to polynomial.

****

AD-NEGF: An End-to-End Differentiable Quantum Transport Simulator for Sensitivity Analysis and Inverse Problems

Zhanghao Zhouyin,Xiang Chen,Peng Zhang,Jun Wang,Lei Wang,Hong Guo

Quantum transport theory describes transport phenomena from first principles, which is essential for domains such as semiconductor fabrication. As a representative, the Non-Equilibrium Green Function (NEGF) method achieves superiority in numerical accuracy. However, its tremendous computational cost makes it unbearable for high-throughput simulation tasks such as sensitivity analysis, inverse design, etc. In this work, we propose AD-NEGF, to the best of our knowledge the first Automatic Differentiation (AD) based quantum transport simulator. AD-NEGF calculates gradient information efficiently by utilizing automatic differentiation and implicit layer techniques, while guaranteeing the correctness of the forward simulation. Such gradient information enables accurate and efficient calculation of differential physical quantities and solving inverse problems that are intractable by traditional optimization methods.

****

Incomplete to complete multiphysics forecasting - a hybrid approach for learning unknown phenomena

Nilam Nandkishor Tathawadekar,Nguyen Anh Khoa Doan,Camilo Fernando Silva,Nils Thuerey

Modeling complex dynamical systems where only partial knowledge of their physical mechanisms is available is a crucial problem across all scientific and engineering disciplines. Purely data-driven approaches, which only make use of an artificial neural network and data, often fail to accurately simulate the evolution of the system dynamics over a sufficiently long time and in a physically consistent manner. Therefore, we propose a hybrid approach that uses a neural network model in combination with an incomplete PDE solver that provides known but incomplete physical information. In this study, we demonstrate that the results obtained from the incomplete PDEs can be efficiently corrected at every time step by the proposed hybrid neural network – PDE solver model, so that the effect of the unknown physics present in the system is correctly accounted for. For validation purposes, the obtained simulations of the hybrid model are successfully compared against results coming from the complete set of PDEs describing the full physics of the considered system. We demonstrate the validity of the proposed approach on a reactive flow, an archetypal multi-physics system that combines fluid mechanics and chemistry, the latter being the physics considered unknown. Experiments are made on planar and Bunsen-type flames at various operating conditions. The hybrid neural network - PDE approach correctly models the flame evolution of the cases under study for significantly long time windows, yields improved generalization, and allows for larger simulation time steps.

****

Bi-Level Dynamic Parameter Sharing among Individuals and Teams for Promoting Collaborations in Multi-Agent Reinforcement Learning

Yan Liu,Ying Tiffany He,Fei Richard Yu,Zhong Ming

Parameter sharing has greatly contributed to the success of multi-agent reinforcement learning in recent years. However, most existing parameter sharing mechanisms are static, and parameters are indiscriminately shared among individuals, ignoring the dynamic environments and different roles of multiple agents. In addition, although a single-level selective parameter sharing mechanism can promote the diversity of strategies, it is hard to establish complementary and cooperativ e relationships between agents. To address these issues, we propose a bi-level dynamic parameter sharing mechanism among individuals and teams for promoting effective collaborations (BDPS). Specifically, at the individual level, we define virtual dynamic roles based on the long-term cumulative advantages of agents and share parameters among agents in the same role. At the team level, we combine agents of different virtual roles and share parameters of agents in the same group. Through the joint efforts of these two levels, we achieve a dynamic balance between the individuality and commonality of agents, enabling agents to learn more complex and complementary collaborative relationships. We evaluate BDPS on a challenging set of StarCraft II micromanagement tasks. The experimental results show that our method outperforms the current state-of-the-art baselines, and we demonstrate the reliability of our proposed structure through ablation experiments.

****

How to prepare your task head for finetuning

Yi Ren,Shangmin Guo,Wonho Bae,Danica J. Sutherland

In the era of deep learning, transferring information from a pretrained network to a downstream task by finetuning has many benefits. The choice of task head plays an important role in fine-tuning, as the pretrained and downstream tasks are usually different. Although there exist many different designs for finetuning, a full understanding of when and why these algorithms work has been elusive. We analyze how the choice of task head controls feature adaptation and hence influences the downstream performance. By decomposing the feature's learning dynamics, we find the key aspect is the training accuracy and loss at the beginning of finetuning, which determines the "energy" available for the feature's adaptation. We identify a significant trend in the effect of changes in this initial energy on the resulting features after finetuning. Specifically, as the energy increases, the Euclidean and cosine distances between the resulting and original features increase, while their dot product (and the resulting features’ norm) first increases and then decreases. Inspired by this, we give several practical principles that lead to better downstream performance. We analytically prove this trend in an overparamterized linear setting and verify its applicability to different experimental settings.

****

Loss Landscapes are All You Need: Neural Network Generalization Can Be Explained Without the Implicit Bias of Gradient Descent

Ping-yeh Chiang,Renkun Ni,David Yu Miller,Arpit Bansal,Jonas Geiping,Micah Goldblum,Tom Goldstein

It is commonly believed that the implicit regularization of optimizers is needed for neural networks to generalize in the overparameterized regime. In this paper, we observe experimentally that this implicit regularization behavior is {\em generic}, i.e. it does not depend strongly on the choice of optimizer. We demonstrate this by training neural networks using several gradient-free optimizers, which do not benefit from properties that are often attributed to gradient-based optimizers.   This includes a guess-and-check optimizer that generates uniformly random parameter vectors until finding one that happens to achieve perfect train accuracy, and a zeroth-order Pattern Search optimizer that uses no gradient computations. In the low sample and few-shot regimes, where zeroth order optimizers are most computationally tractable, we find that these non-gradient optimizers achieve test accuracy comparable to SGD. The code to reproduce results can be found at https://github.com/Ping-C/optimizer .

****

DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models

Shansan Gong,Mukai Li,Jiangtao Feng,Zhiyong Wu,Lingpeng Kong

Recently, diffusion models have emerged as a new paradigm for generative models. Despite the success in domains using continuous signals such as vision and audio, adapting diffusion models to natural language is under-explored due to the discrete nature of texts, especially for conditional generation. We tackle this challenge by proposing DiffuSeq: a diffusion model designed for sequence-to-sequence (Seq2Seq) text generation tasks. Upon extensive evaluation over a wide range of Seq2Seq tasks, we find DiffuSeq achieving comparable or even better performance than six established baselines, including a state-of-the-art model that is based on pre-trained language models. Apart from quality, an intriguing property of DiffuSeq is its high diversity during generation, which is desired in many Seq2Seq tasks. We further include a theoretical analysis revealing the connection between DiffuSeq and autoregressive/non-autoregressive models. Bringing together theoretical analysis and empirical evidence, we demonstrate the great potential of diffusion models in complex conditional language generation tasks. Code is available at https://github.com/Shark-NLP/DiffuSeq

****

Rethinking Deep Spiking Neural Networks: A Multi-Layer Perceptron Approach

Luziwei Leng,Boyan Li,Ran Cheng,Shuaijie Shen,Kaixuan Zhang,Jianguo Zhang,Jianxing Liao

By adopting deep convolution architectures, spiking neural networks (SNNs) have recently achieved competitive performances with their artificial counterparts in image classification, meanwhile with much lower computation cost due to event-driven and sparse activation. However, the multiplication-free inference (MFI) principle makes SNNs incompatible with attention or transformer mechanisms which have shown significant performance gains on high resolution vision tasks. Inspired from recent works on multi-layer perceptrons (MLPs), we explore an efficient spiking MLP design using batch normalization instead of layer normalization in both the token and the channel block to be compatible with MFI. We further strengthen the network’s local feature learning ability with a spiking patch encoding layer, which significantly improves the network performance. Based on these building blocks, we explore an optimal skip connection configuration and develop an efficient multi-stage spiking MLP network combining global receptive field and local feature extraction, achieving full spike-based computation. Without pre-training or other advanced SNN training techniques, the spiking MLP network achieves 66.39% top-1 accuracy on the ImageNet-1K dataset, surpassing the state-of-the-art directly trained spiking ResNet-34 by 2.67% under similar model capacity meanwhile with shorter simulation steps and much less computation cost. Another larger variant of the network achieves 68.84% top-1 accuracy, rivaling the spiking VGG-16 network with 4 times smaller model capacity. Our work demonstrates the effectiveness of an alternative deep SNN architecture combining both global and local learning abilities. More interestingly, finally we show a close resemblance of the trained receptive field of our network to cells in the cortex. Code will be publicly available.

****

Collaborative Symmetricity Exploitation for Offline Learning of Hardware Design Solver

Haeyeon Kim,Minsu Kim,Joungho Kim,Jinkyoo Park

This paper proposes \textit{collaborative symmetricity exploitation} (CSE) framework to train a solver for the decoupling capacitor placement problem (DPP) benchmark, one of the significant hardware design problems. Due to the sequentially coupled multi-level property of the hardware design process, the design condition of DPP changes depending on the design of higher-level problems. Also, the online evaluation of real-world electrical performance through simulation is extremely costly. Thus, a data-efficient offline learning method to train a solver (i.e., contextualized policy) with high generalization capability over changing task conditions is necessary. In this paper, we apply the CSE framework to train a DPP solver using a limited number of offline expert data. Leveraging the symmetricity for offline learning of hardware design solver has two major advantages: it increases data-efficiency by reducing the solution space and improves generalization capability by capturing the invariant nature present regardless of changing conditions. The proposed CSE is composed of two learning schemes: expert exploitation and self-exploitation. Expert exploitation induces symmetricity during the imitation learning process with offline expert data and self-exploitation induces symmetricity during the consistency learning process with self-generated data. Extensive experiments verified that CSE with zero-shot inference outperforms the neural baselines and iterative conventional design methods on the DPP benc hmark. Furthermore, CSE showed promising extrapolation capability as it greatly outperforms the expert method used to generate the offline data for training. Scalability and flexibility of the proposed method were also verified for practical use of CSE in industry.

****

Policy Expansion for Bridging Offline-to-Online Reinforcement Learning

Haichao Zhang,Wei Xu,Haonan Yu

Pre-training with offline data and online fine-tuning using reinforcement learning is a promising strategy for learning control policies by leveraging the best of both worlds in terms of sample efficiency and performance. One natural approach is to initialize the policy for online learning with the one trained offline. In this work, we introduce a policy expansion scheme for this task. After learning the offline policy, we use it as one candidate policy in a policy set, and further learn another policy that will be responsible for further learning as an expansion to the policy set. The two policies will be composed in an adaptive manner for interacting with the environment. With this approach, the policy previously learned offline is fully retained during online learning, thus mitigating the potential issues such as destroying the useful behaviors of the offline policy in the initial stage of online learning while allowing the offline policy participate in the exploration naturally in an adaptive manner. Moreover, new useful behaviors can potentially be captured by the newly added policy through learning.

Experiments are conducted on a number of tasks and the results demonstrate the effectiveness of the proposed approach.

****

On The Implicit Bias of Weight Decay in Shallow Univariate ReLU Networks

Boris Hanin

We give a complete characterization of the implicit bias of infinitesimal weight decay in the modest setting of univariate one layer ReLU networks. Our main result is a surprisingly simple geometric description of all one layer ReLU networks that exactly fit a dataset $\mathcal D= \set{(x_i,y_i)}$ with the minimum value of the ℓ_2-norm of the neuron weights. Specifically, we prove that such functions must be either concave or convex between any two consecutive data sites x_i and x_{i+1}. Our description implies that interpolating ReLU networks with weak ℓ_2-regularization achieve the best possible generalization for learning $1d$ Lipschitz functions, up to universal constants.

****

Mitigating Memorization of Noisy Labels via Regularization between Representations

Hao Cheng,Zhaowei Zhu,Xing Sun,Yang Liu

Designing robust loss functions is popular in learning with noisy labels while existing designs did not explicitly consider the overfitting property of deep neural networks (DNNs). As a result, applying these losses may still suffer from overfitting/memorizing noisy labels as training proceeds. In this paper, we first theoretically analyze the memorization effect and show that a lower-capacity model may perform better on noisy datasets. However, it is non-trivial to design a neural network with the best capacity given an arbitrary task. To circumvent this dilemma, instead of changing the model architecture, we decouple DNNs into an encoder followed by a linear classifier and propose to restrict the function space of a DNN by a representation regularizer. Particularly, we require the distance between two self-supervised features to be positively related to the distance between the corresponding two supervised model outputs.

Our proposed framework is easily extendable and can incorporate many other robust loss functions to further improve performance. Extensive experiments and theoretical analyses support our claims. Code is available at https://github.com/UCSC-REAL/SelfSup_NoisyLabel.

****

Graph Neural Networks are Inherently Good Generalizers: Insights by Bridging GNNs and MLPs

Chenxiao Yang,Qitian Wu,Jiahua Wang,Junchi Yan

Graph neural networks (GNNs), as the de-facto model class for representation learning on graphs, are built upon the multi-layer perceptrons (MLP) architecture with additional message passing layers to allow features to flow across nodes. While conventional wisdom commonly attributes the success of GNNs to their advanced expressivity, we conjecture that this is not the main cause of GNNs' superiority in node-level prediction tasks. This paper pinpoints the major source of GNNs' performance gain to their intrinsic generalization capability, by introducing an intermediate model class dubbed as P(ropagational)MLP, which is identical to standard MLP in training, but then adopts GNN's architecture in testing. Intriguingly, we observe that PMLPs consistently perform on par with (or even exceed) their GNN counterparts, while being much more efficient in training.

This finding provides a new perspective for understanding the learning behavior of GNNs, and can be used as an analytic tool for dissecting various GNN-related research problems including expressivity, generalization, over-smoothing and heterophily. As an initial step to analyze PMLP, we show its essential difference to MLP at infinite-width limit lies in the NTK feature map in the post-training stage. Moreover, through extrapolation analysis (i.e., generalization under distribution shifts), we find that though most GNNs and their PMLP counterparts cannot extrapolate non-linear functions for extreme out-of-distribution data, they have greater potential to generalize to testing data near the training data support as natural advantages of the GNN architecture used for inference.
****

Learning Cut Selection for Mixed-Integer Linear Programming via Hierarchical Sequence Model

Zhihai Wang,Xijun Li,Jie Wang,Yufei Kuang,Mingxuan Yuan,Jia Zeng,Yongdong Zhang,Feng Wu

Cutting planes (cuts) are important for solving mixed-integer linear programs (MILPs), which formulate a wide range of important real-world applications. Cut selection---which aims to select a proper subset of the candidate cuts to improve the efficiency of solving MILPs---heavily depends on (P1) which cuts should be preferred, and (P2) how many cuts should be selected. Although many modern MILP solvers tackle (P1)-(P2) by manually designed heuristics, machine learning offers a promising approach to learn more effective heuristics from MILPs collected from specific applications. However, many existing learning-based methods focus on learning which cuts should be preferred, neglecting the importance of learning the number of cuts that should be selected. Moreover, we observe from extensive empirical results that (P3) what order of selected cuts should be preferred has a significant impact on the efficiency of solving MILPs as well. To address this challenge, we propose a novel hierarchical sequence model (HEM) to learn cut selection policies via reinforcement learning. Specifically, HEM consists of a two-level model: (1) a higher-level model to learn the number of cuts that should be selected, (2) and a lower-level model---that formulates the cut selection task as a sequence to sequence learning problem---to learn policies selecting an ordered subset with the size determined by the higher-level model. To the best of our knowledge, HEM is the first method that can tackle (P1)-(P3) in cut selection simultaneously from a data-driven perspective. Experiments show that HEM significantly improves the efficiency of solving MILPs compared to human-designed and learning-based baselines on both synthetic and large-scale real-world MILPs, including MIPLIB 2017. Moreover, experiments demonstrate that HEM well generalizes to MILPs that are significantly larger than those seen during training.
****

BSTT: A Bayesian Spatial-Temporal Transformer for Sleep Staging

Yuchen Liu,Ziyu Jia

Sleep staging is helpful in assessing sleep quality and diagnosing sleep disorders. However, how to adequately capture the temporal and spatial relations of the brain during sleep remains a challenge. In particular, existing methods cannot adaptively infer spatial-temporal relations of the brain under different sleep stages. In this paper, we propose a novel Bayesian spatial-temporal relation inference neural network, named Bayesian spatial-temporal transformer (BSTT), for sl eep staging. Our model is able to adaptively infer brain spatial-temporal relations during sleep for spatial-temporal feature modeling through a well-designed Bayesian relation inference component. Meanwhile, our model also includes a spatial transformer for extracting brain spatial features and a temporal transformer for capturing temporal features. Experiments show that our BSTT outperforms state-of-the-art baselines on ISRUC and MASS datasets. In addition, the visual analysis shows that the spatial-temporal relations obtained by BSTT inference have certain interpretability for sleep staging.

****

Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning

Jiahui Gao,Renjie Pi,LIN Yong,Hang Xu,Jiacheng Ye,Zhiyong Wu,WEIZHONG ZHANG,Xiaodan Liang,Zhenguo Li,Lingpeng Kong

There is a rising interest in further exploring the zero-shot learning potential of large pre-trained language models (PLMs). A new paradigm called data-generation-based zero-shot learning has achieved impressive success. In this paradigm, the synthesized data from the PLM acts as the carrier of knowledge, which is used to train a task-specific model with orders of magnitude fewer parameters than the PLM, achieving both higher performance and efficiency than prompt-based zero-shot learning methods on PLMs. The main hurdle of this approach is that the synthesized data from PLM usually contains a significant portion of low-quality samples. Fitting on such data will greatly hamper the performance of the task-specific model, making it unreliable for deployment. Previous methods remedy this issue mainly by filtering synthetic data using heuristic metrics(e.g., output confidence), or refining the data with the help of a human expert, which comes with excessive manual tuning or expensive costs. In this paper, we propose a novel noise-robust re-weighting framework SunGen to automatically construct high-quality data for zero-shot classification problems. Our framework features the ability to learn the sample weights indicating data quality without requiring any human annotation. We theoretically and empirically verify the ability of our method to help construct good-quality synthetic datasets. Notably, SunGen-LSTM yields a 9.8% relative improvement than the baseline on average accuracy across eight different established text classification tasks.

****

Beyond re-balancing: distributionally robust augmentation against class-conditional distribution shift in long-tailed recognition

Keliang Li,Hong Chang,Shiguang Shan,Xilin CHEN

As a fundamental and practical problem, long-tailed recognition has drawn burning attention. In this paper, we investigate an essential but rarely noticed issue in long-tailed recognition, Class-Conditional Distribution (CCD) shift due to scarce instances, which exhibits a significant discrepancy between the empirical CCDs for training and test data, especially for tail classes. We observe empirical evidence that the shift is a key factor that limits the performance of existing long-tailed learning methods, and provide novel understanding of these methods in the
course of our analysis. Motivated by this, we propose an adaptive data augmentation method, Distributionally Robust Augmentation (DRA), to learn models more robust to CCD shift. The convergence and generalization of DRA are theoretically guaranteed. Experimental results verify that DRA outperforms related data augmentation methods without extra training cost and significantly improves the performance of some existing long-tailed recognition methods.

****

Improving Deep Policy Gradients with Value Function Search

Enrico Marchesini,Christopher Amato

Deep Policy Gradient (PG) algorithms employ value networks to drive the learning of parameterized policies and reduce the variance of the gradient estimates. However, value function approximation gets stuck in local optima and struggles to fit the actual return, limiting the variance reduction efficacy and leading policies to sub-optimal performance. This paper focuses on improving value approximation and analyzing the effects on Deep PG primitives such as value prediction, v ariance reduction, and correlation of gradient estimates with the true gradient. To this end, we introduce a Value Function Search that employs a population of perturbed value networks to search for a better approximation. Our framework does not require additional environment interactions, gradient computations, or ensembles, providing a computationally inexpensive approach to enhance the supervised learning task on which value networks train. Crucially, we show that improving Deep PG primitives results in improved sample efficiency and policies with higher returns using common continuous control benchmark domains.

****

MEDICAL IMAGE UNDERSTANDING WITH PRETRAINED VISION LANGUAGE MODELS: A COMPREHENSIVE STUDY

Ziyuan Qin,Hua Hui Yi,Qicheng Lao,Kang Li

The large-scale pre-trained vision language models (VLM) have shown remarkable domain transfer capability on natural images. However, it remains unknown whether this capability can also apply to the medical image domain. This paper thoroughly studies the knowledge transferability of pre-trained VLMs to the medical domain, where we show that well-designed medical prompts are the key to elicit knowledge from pre-trained VLMs. We demonstrate that by prompting with expressive attributes that are shared between domains, the VLM can carry the knowledge across domains and improve its generalization. This mechanism empowers VLMs to recognize novel objects with fewer or without image samples. Furthermore, to avoid the laborious manual designing process, we develop three approaches for automatic generation of medical prompts, which can inject expert-level medical knowledge and image-specific information into the prompts for fine-grained grounding. We conduct extensive experiments on thirteen different medical datasets across various modalities, showing that our well-designed prompts greatly improve the zero-shot performance compared to the default prompts, and our fine-tuned models surpass the supervised models by a significant margin.

****

Temporal Coherent Test Time Optimization for Robust Video Classification

Chenyu Yi,SIYUAN YANG,Yufei Wang,Haoliang Li,Yap-peng Tan,Alex Kot

Deep neural networks are likely to fail when the test data is corrupted in real-world deployment (e.g., blur, weather, etc.). Test-time optimization is an effective way that adapts models to generalize to corrupted data during testing, which has been shown in the image domain. However, the techniques for improving video classification corruption robustness remain few. In this work, we propose a Temporal Coherent Test-time Optimization framework (TeCo) to utilize spatio-temporal information in test-time optimization for robust video classification. To exploit information in video with self-supervised learning, TeCo minimizes the entropy of the prediction based on the global content from video clips. Meanwhile, it also feeds local content to regularize the temporal coherence at the feature level. TeCo retains the generalization ability of various video classification models and achieves significant improvements in corruption robustness across Mini Kinetics-C and Mini SSV2-C. Furthermore, TeCo sets a new baseline in video classification corruption robustness via test-time optimization.

****

Offline Communication Learning with Multi-source Datasets

Yihuan Mao,Rui Hu,Lulu Zheng,Jianhao Wang,Chongjie Zhang

Scalability and partial observability are two major challenges faced by multi-agent reinforcement learning. Recently researchers propose offline MARL algorithms to improve scalability by reducing online exploration cost, while the problem of partial observability is often ignored in the offline MARL setting. Communication is a promising approach to alleviate the miscoordination caused by partially observability, thus in this paper we focus on offline communication learning where agents learn from an fixed dataset. We find out that learning communications in an end-to-end manner from a given offline dateset without communication information is intractable, since the correct communication protocol space is too sparse compared with the exponentially growing joint state-action space when the number of agents increases. Besides, unlike offline policy learning which can be guided by reward signals, offline communication learning is struggling since com munication messages implicitly impact the reward. Moreover, in real-world applications, offline MARL datasets are often collected from multi-source, leaving offline MARL communication learning more challenging. Therefore, we present a new benchmark which contains a diverse set of challenging offline MARL communication tasks with single/multi-source datasets, and propose a novel Multi-Head structure for Communication Imitation learning (MHCI) algorithm that automatically adapts to the distribution of the dataset. Empirical result shows the effectiveness of our method on various tasks of the new offline communication learning benchmark.

****

A Learning Based Hypothesis Test for Harmful Covariate Shift

Tom Ginsberg,Zhongyuan Liang,Rahul G Krishnan

The ability to quickly and accurately identify covariate shift at test time is a critical and often overlooked component of safe machine learning systems deployed in high-risk domains. While methods exist for detecting when predictions should not be made on out-of-distribution test examples, identifying distributional level differences between training and test time can help determine when a model should be removed from the deployment setting and retrained. In this work, we define harmful covariate shift (HCS) as a change in distribution that may weaken the generalization of a predictive model. To detect HCS, we use the discordance between an ensemble of classifiers trained to agree on training data and disagree on test data. We derive a loss function for training this ensemble and show that the disagreement rate and entropy represent powerful discriminative statistics for HCS. Empirically, we demonstrate the ability of our method to detect harmful covariate shift with statistical certainty on a variety of high-dimensional datasets. Across numerous domains and modalities, we show state-of-the-art performance compared to existing methods, particularly when the number of observed test samples is small.

****

Less is More: Rethinking Few-Shot Learning and Recurrent Neural Nets

Deborah Pereg,Martin Villiger,Brett Bouma,Polina Golland

The statistical supervised learning framework assumes an input-output set with a joint probability distribution that is reliably represented by the training dataset. The learner is then required to output a prediction rule learned from the training dataset's input-output pairs. In this work, we provide meaningful insights into the asymptotic equipartition property (AEP) \citep{Shannon:1948} in the context of machine learning, and illuminate some of its potential ramifications for few-shot learning. We provide theoretical guarantees for reliable learning under the information-theoretic AEP, and for the generalization error with respect to the sample size. We then focus on a highly efficient recurrent neural net (RNN) framework and propose a reduced-entropy algorithm for few-shot learning. We also propose a mathematical intuition for the RNN as an approximation of a sparse coding solver. We verify the applicability, robustness, and computational efficiency of the proposed approach with image deblurring and optical coherence tomography (OCT) speckle suppression. Our experimental results demonstrate significant potential for improving learning models' sample efficiency, generalization, and time complexity, that can therefore be leveraged for practical real-time applications.

****

Deep Transformers without Shortcuts: Modifying Self-attention for Faithful Signal Propagation

Bobby He,James Martens,Guodong Zhang,Aleksandar Botev,Andrew Brock,Samuel L Smith,Yee Whye Teh

Skip connections and normalisation layers form two standard architectural components that are ubiquitous for the training of Deep Neural Networks (DNNs), but whose precise roles are poorly understood. Recent approaches such as Deep Kernel Shaping have made progress towards reducing our reliance on them, using insights from wide NN kernel theory to improve signal propagation in vanilla DNNs (which we define as networks without skips or normalisation). However, these approaches are incompatible with the self-attention layers present in transformers, whose

kernels are intrinsically more complicated to analyse and control. And so the question remains: \emph{is it possible to train deep vanilla transformers?} We answer this question in the affirmative by designing several approaches that use combinations of parameter initialisations, bias matrices and location-dependent rescaling to achieve faithful signal propagation in vanilla transformers. Our methods address various intricacies specific to signal propagation in transformers, including the interaction with positional encoding and causal masking. In experiments on WikiText-103 and C4, our approaches enable deep transformers without normalisation to train at speeds matching their standard counterparts, and deep vanilla transformers to reach the same performance as standard ones after about 5 times more iterations.
****
Towards Understanding Robust Memorization in Adversarial Training
Binghui Li,Yuanzhi Li
Adversarial training is a standard method to train neural networks to be robust to adversarial perturbation. However, in contrast with benign overfitting in the standard deep learning setting, which means that over-parameterized neural networks surprisingly generalize well for unseen data, while adversarial training method is able to achieve low robust training error, there still exists a significant robust generalization gap, which promotes us exploring what mechanism leads to robust overfitting during learning process. In this paper, we propose an implicit bias called $\textit{robust memorization}$ in adversarial training under the realistic data assumption. By function approximation theory, we prove that ReLU nets with efficient size have the ability to achieve robust memorization, while robust generalization requires exponentially large models. Then, we demonstrate robust memorization in adversarial training from both empirical and theoretical perspectives. In particular, we empirically investigate the dynamics of loss landscape over input, and we also provide theoretical analysis of robust memorization on data with linear separable assumption. Finally, we prove novel generalization bounds based on robust memorization, which further explains why deep neural networks have both high clean test accuracy and robust overfitting at the same time.
****
Self-Supervised Geometric Correspondence for Category-Level 6D Object Pose Estimation in the Wild
Kaifeng Zhang,Yang Fu,Shubhankar Borse,Hong Cai,Fatih Porikli,Xiaolong Wang
While 6D object pose estimation has wide applications across computer vision and robotics, it remains far from being solved due to the lack of annotations. The problem becomes even more challenging when moving to category-level 6D pose, which requires generalization to unseen instances. Current approaches are restricted by leveraging annotations from simulation or collected from humans. In this paper, we overcome this barrier by introducing a self-supervised learning approach trained directly on large-scale real-world object videos for category-level 6D pose estimation in the wild. Our framework reconstructs the canonical 3D shape of an object category and learns dense correspondences between input images and the canonical shape via surface embedding. For training, we propose novel geometrical cycle-consistency losses which construct cycles across 2D-3D spaces, across different instances and different time steps. The learned correspondence can be applied for 6D pose estimation and other downstream tasks such as keypoint transfer. Surprisingly, our method, without any human annotations or simulators, can achieve on-par or even better performance than previous supervised or semi-supervised methods on in-the-wild images. Code and videos are available at https://kywind.github.io/self-pose.
****
Incorporating Explicit Uncertainty Estimates into Deep Offline Reinforcement Learning
David Brandfonbrener,Remi Tachet des Combes,Romain Laroche
Most theoretically motivated work in the offline reinforcement learning setting requires precise uncertainty estimates. This requirement restricts the algorithms derived in that work to the tabular and linear settings where such estimates e

xist. In this work, we develop a novel method for incorporating scalable uncertainty estimates into an offline reinforcement learning algorithm called deep-SPIBB that extends the SPIBB family of algorithms to environments with larger state and action spaces. We use recent innovations in uncertainty estimation from the deep learning community to get more scalable uncertainty estimates to plug into deep-SPIBB. While these uncertainty estimates do not allow for the same theoretical guarantees as in the tabular case, we argue that the SPIBB mechanism for incorporating uncertainty is more robust and flexible than pessimistic approaches that incorporate the uncertainty as a value function penalty. We bear this out empirically, showing that deep-SPIBB outperforms pessimism based approaches with access to the same uncertainty estimates and performs at least on par with a variety of other strong baselines across several environments and datasets.

****

Non-parametric Outlier Synthesis

Leitian Tao,Xuefeng Du,Jerry Zhu,Yixuan Li

Out-of-distribution (OOD) detection is indispensable for safely deploying machine learning models in the wild. One of the key challenges is that models lack supervision signals from unknown data, and as a result, can produce overconfident predictions on OOD data. Recent work on outlier synthesis modeled the feature space as parametric Gaussian distribution, a strong and restrictive assumption that might not hold in reality. In this paper, we propose a novel framework, non-parametric outlier synthesis (NPOS), which generates artificial OOD training data and facilitates learning a reliable decision boundary between ID and OOD data. Importantly, our proposed synthesis approach does not make any distributional assumption on the ID embeddings, thereby offering strong flexibility and generality. We show that our synthesis approach can be mathematically interpreted as a rejection sampling framework. Extensive experiments show that NPOS can achieve superior OOD detection performance, outperforming the competitive rivals by a significant margin. Code is publicly available at https://github.com/deeplearning-wisc/npos.

****

Robust Self-Supervised Learning with Lie Groups

Mark Ibrahim,Diane Bouchacourt,Ari S. Morcos

Deep learning has led to remarkable advances in computer vision. Even so, today’s best models are brittle when presented with variations that differ even slightly from those seen during training. Minor shifts in the pose, color, or illumination of an object can lead to catastrophic misclassifications. State-of-the art models struggle to understand how a set of variations can affect different objects. We propose a framework for instilling a notion of how objects vary in more realistic settings. Our approach applies the formalism of Lie groups to capture continuous transformations to improve models’ robustness to distributional shifts. We apply our framework on top of state-of-the-art self-supervised learning (SSL) models, finding that explicitly modeling transformations with Lie groups leads to substantial performance gains of greater than 10% for MAE on both known instances seen in typical poses now presented in new poses, and on unknown instances in any pose. We also apply our approach to ImageNet, finding that the Lie operator improves performance by almost 4%. These results demonstrate the promise of learning transformations to improve model robustness.

****

Self-Paced Learning Enhanced Physics-informed Neural Networks for Solving Partial Differential Equations

Xiaoting Han

There is a hit discussion on solving partial differential equation by neural network. The famous PINN (physics-informed neural networks) has drawn worldwide attention since it was put forward. Despite its success in solving nonlinear partial differential equation, the difficulty in converging and the inefficiency in training process are definitely huge concerns. Normally, data for PINN is randomly chosen for a given distribution. Additionally, it's fitted to a model in a meaningless way. Curriculum Learning is a learning strategy that trains a model from easy samples to hard ones, which represents the meaningful human learning order

. Self-paced Learning (SPL) is one of the significant branches of Automatic Curriculum Learning, which takes example-wise the training loss as Difficulty Measurer. SPL is an efficient strategy in enhancing the convergence rate of numerous models. In this paper, we propose a novel SPL-PINN learning framework, with SPL to accelerate the convergence progress of PINN. We demonstrate the effectiveness of SPL-PINN in a typical parabolic equation and Burgers equation.

****

Moving Beyond Handcrafted Architectures in Self-Supervised Learning

Sharath Girish,Debadeepta Dey,Neel Joshi,Vibhav Vineet,Shital Shah,Caio Cesar Teodoro Mendes,Abhinav Shrivastava,Yale Song

The current literature on self-supervised learning (SSL) focuses on developing learning objectives to train neural networks more effectively on unlabeled data. The typical development process involves taking well-established architectures, e.g., ResNet demonstrated on ImageNet, and using them to evaluate newly developed objectives on downstream scenarios. While convenient, this does not take into account the role of architectures which has been shown to be crucial in the supervised learning literature. In this work, we establish extensive evidence showing that architecture plays a significant role in SSL. We conduct a large-scale study with over 100 variants of ResNet and MobileNet architectures and evaluate them across 11 downstream scenarios in the SSL setting. We show that there is no one network that performs consistently well across the scenarios. Based on this, we propose to learn not only network weights but also architecture topologies in the SSL regime. We show that ``self-supervised architectures'' significantly outperform popular handcrafted architectures (ResNet-50 and MobileNetV2) on major image classification benchmarks (ImageNet-1K, iNat2021, and more). Our results suggest that it is time to consider moving beyond handcrafted architectures in SSL and start thinking about incorporating architecture search into self-supervised learning objectives.

****

Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks

Namjoon Suh,Tian-Yi Zhou,Xiaoming Huo

We develop a new approximation and estimation analysis of deep feed-forward neural networks (FNNs) with the Rectified Linear Unit (ReLU) activation. The functions of interests for the approximation and estimation are assumed to be from Sobolev spaces defined over the d-dimensional unit sphere with smoothness index $r>0$. In the regime where r is in the constant order (i.e., $r=\mathcal{O}(1)$), it is shown that at most d^d active parameters are required for getting d^{-C} approximation rate for some constant $C>0$. In contrast, in the regime where the index r grows in the order of d (i.e., $r=\mathcal{O}(d)$) asymptotically, we prove the approximation error decays in the rate $d^{-d^{\beta}}$ with $0<\beta<1$ up to some constant factor independent of d. The required number of active parameters in the networks for the approximation increases polynomially in d as $d\rightarrow{\infty}$. In addition to this, it is shown that bound on the excess risk has a d^d factor, when $r=\mathcal{O}(1)$, whereas it has $d^{\mathcal{O}(1)}$ factor, when $r=\mathcal{O}(d)$. We emphasize our findings by making comparisons to the results on approximation and estimation errors of deep ReLU FNN when functions are from Sobolev spaces defined over d-dimensional cube. Here, we show that with the current state-of-the-art result, d^{d} factor remain both in the approximation and estimation error, regardless of the order of r.

****

Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement

Chongyi Li,Chun-Le Guo,man zhou,Zhexin Liang,Shangchen Zhou,Ruicheng Feng,Chen Change Loy

Ultra-High-Definition (UHD) photo has gradually become the standard configuration in advanced imaging devices. The new standard unveils many issues in existing approaches for low-light image enhancement (LLIE), especially in dealing with the intricate issue of joint luminance enhancement and noise removal while remaining efficient. Unlike existing methods that address the problem in the spatial do main, we propose a new solution, UHDFour, that embeds Fourier transform into a cascaded network. Our approach is motivated by a few unique characteristics in the Fourier domain: 1) most luminance information concentrates on amplitudes while noise is closely related to phases, and 2) a high-resolution image and its low-resolution version share similar amplitude patterns. Through embedding Fourier into our network, the amplitude and phase of a low-light image are separately processed to avoid amplifying noise when enhancing luminance. Besides, UHDFour is scalable to UHD images by implementing amplitude and phase enhancement under the low-resolution regime and then adjusting the high-resolution scale with few computations. We also contribute the first real UHD LLIE dataset, UHD-LL, that contains 2,150 low-noise/normal-clear 4K image pairs with diverse darkness and noise levels captured in different scenarios. With this dataset, we systematically analyze the performance of existing LLIE methods for processing UHD images and demonstrate the advantage of our solution. We believe our new framework, coupled with the dataset, would push the frontier of LLIE towards UHD. The code and dataset are available at https://li-chongyi.github.io/UHDFour/.

****

Population-Based Reinforcement Learning for Combinatorial Optimization Problems

Nathan Grinsztajn,Daniel Furelos-Blanco,Thomas D Barrett

Applying reinforcement learning to combinatorial optimization problems is attractive as it obviates the need for expert knowledge or pre-solved instances. However, it is unrealistic to expect an agent to solve these (often NP-)hard problems in a single shot at inference due to their inherent complexity, thus leading approaches are often augmented with additional search strategies, from stochastic sampling and beam-search to explicit fine-tuning.

In this paper, we argue for the benefits of learning a population of complementary agents, which can be simultaneously rolled out at inference. To this end, we introduce Poppy, a simple theoretically grounded training procedure for populations. Instead of relying on a predefined or hand-crafted notion of diversity, Poppy induces an unsupervised specialization targeted solely at maximizing the performance of the whole population. We show that Poppy leads to a set of complementary heuristics, and obtain state-of-the-art results on three popular NP-hard problems: the traveling salesman (TSP), the capacitated vehicle routing (CVRP), and 0-1 knapsack (KP). On TSP specifically, Poppy divides by 5 the optimality gap while reducing the inference time by more than 10 compared to previous state-of-the-art reinforcement learning approaches.

****

Adversarial Attack Detection Through Network Transport Dynamics

Skander Karkar,patrick gallinari,Alain Rakotomamonjy

Adversarial attacks are perturbations to the input that don't change its class for a human observer, but fool a neural network into changing its prediction. In this paper, we propose a detector of such attacks that is based on the view of residual networks as discrete dynamical systems. The detector tells clean inputs from abnormal ones by comparing the discrete vector fields they follow throughout the network's layers before the final classification layer. We compare this detector favorably to other detectors on seen and unseen attacks. We also show that regularizing this vector field during training makes the network more regular on the data distribution's support, thus making the network's activations on clean samples more distinguishable from those of abnormal samples. This regularization of the network's dynamics improves the performance of any detection method that uses the internal embeddings as inputs, while also improving the network's test accuracy.

****

On the Relationship Between Adversarial Robustness and Decision Region in Deep Neural Networks

Seongjin Park,Haedong Jeong,Giyoung Jeon,Jaesik Choi

In general, Deep Neural Networks (DNNs) are evaluated by the generalization performance measured on unseen data excluded from the training phase. Along with the development of DNNs, the generalization performance converges to the state-of-the-art and it becomes difficult to evaluate DNNs solely based on this metric. Th e robustness against adversarial attack has been used as an additional metric to evaluate DNNs by measuring their vulnerability. However, few studies have been performed to analyze the adversarial robustness in terms of the geometry in DNNs. In this work, we perform an empirical study to analyze the internal properties of DNNs that affect model robustness under adversarial attacks. In particular, we propose the novel concept of the Populated Region Set (PRS), where training samples are populated more frequently, to represent the internal properties of DNNs in a practical setting. From systematic experiments with the proposed concept, we provide empirical evidence to validate that a low PRS ratio has a strong relationship with the adversarial robustness of DNNs. We also devise PRS regularizer leveraging the characteristics of PRS to improve the adversarial robustness without adversarial training.

****

Confounder Identification-free Causal Visual Feature Learning

Xin Li,Zhizheng Zhang,Guoqiang Wei,Cuiling Lan,Wenjun Zeng,Xin Jin,Zhibo Chen

Confounders in deep learning are in general detrimental to model's generalization where they infiltrate feature representations. Therefore, learning causal features that are free of interference from confounders is important. Most previous causal learning based approaches employ back-door criterion to mitigate the adverse effect of certain specific confounder, which require the explicit identification of confounder. However, in real scenarios, confounders are typically diverse and difficult to be identified. In this paper, we propose a novel Confounder Identification-free Causal Visual Feature Learning (CICF) method, which obviates the need for identifying confounders. CICF models the interventions among different samples based on front-door criterion, and then approximates the global-scope intervening effect upon the instance-level interventions from the perspective of optimization. In this way, we aim to find a reliable optimization direction, which avoids the intervening effects of confounders, to learn causal features. Furthermore, we uncover the relation between CICF and the popular meta-learning strategy MAML, and provide an interpretation of why MAML works from the theoretical perspective of causal learning for the first time. Thanks to the effective learning of causal features, our CICF enables models to have superior generalization capability. Extensive experiments on domain generalization benchmark datasets demonstrate the effectiveness of our CICF, which achieves the state-of-the-art performance.

****

Enhanced Temporal Knowledge Embeddings with Contextualized Language Representations

Zhen Han,Ruotong Liao,Beiyan Liu,Yao Zhang,Zifeng Ding,Jindong Gu,Heinz Koeppl,Hinrich Schuetze,Volker Tresp

World knowledge exists in both structured (tables, knowledge graphs) and unstructured forms (texts). Recently, there have been extensive research efforts in the integration of structured factual knowledge and unstructured textual knowledge. However, most studies focus on incorporating static factual knowledge into pre-trained language models, while there is less work on enhancing temporal knowledge graph embedding using textual knowledge. Existing integration approaches can not apply to temporal knowledge graphs (tKGs) since they often assume knowledge embedding is time-invariant. In fact, the entity embedding in tKG embedding models usually evolves over time, which poses the challenge of aligning temporally relevant textual information with entities. To this end, we propose Enhanced Temporal Knowledge Embeddings with Contextualized Language Representations (ECOLA), which uses tKG quadruple as an implicit measure to temporally align textual data and the time-evolving entity representations and uses a novel knowledge-text prediction task to inject textual information into temporal knowledge embedding. ECOLA jointly optimizes the knowledge-text prediction objective and the temporal knowledge embedding objective, and thus, can simultaneously take full advantage of textual and structured knowledge. Since existing datasets do not provide tKGs with aligned textual data, we introduce three new datasets for training and evaluating ECOLA. Experimental results on the temporal knowledge graph completion task show that ECOLA outperforms state-of-the-art tKG embedding models by a large

margin.
****

Learning Adversarial Linear Mixture Markov Decision Processes with Bandit Feedback and Unknown Transition

Canzhe Zhao,Ruofeng Yang,Baoxiang Wang,Shuai Li

We study reinforcement learning (RL) with linear function approximation, unknown transition, and adversarial losses in the bandit feedback setting. Specifically, the unknown transition probability function is a linear mixture model \citep{AyoubJSWY20,ZhouGS21,HeZG22} with a given feature mapping, and the learner only observes the losses of the experienced state-action pairs instead of the whole loss function. We propose an efficient algorithm LSUOB-REPS which achieves $\widetilde{O}(dS^2\sqrt{K}+\sqrt{HSAK})$ regret guarantee with high probability, where d is the ambient dimension of the feature mapping, S is the size of the state space, A is the size of the action space, H is the episode length and K is the number of episodes. Furthermore, we also prove a lower bound of order $\Omega(dH\sqrt{K}+\sqrt{HSAK})$ for this setting. To the best of our knowledge, we make the first step to establish a provably efficient algorithm with a sublinear regret guarantee in this challenging setting and solve the open problem of \citet{HeZG22}.

****

Weakly Supervised Knowledge Transfer with Probabilistic Logical Reasoning for Object Detection

Martijn Oldenhof,Adam Arany,Yves Moreau,Edward De Brouwer

Training object detection models usually requires instance-level annotations, such as the positions and labels of all objects present in each image. Such supervision is unfortunately not always available and, more often, only image-level information is provided, also known as weak supervision.

Recent works have addressed this limitation by leveraging knowledge from a richly annotated domain. However, the scope of weak supervision supported by these approaches has been very restrictive, preventing them to use all available information. In this work, we propose ProbKT, a framework based on probabilistic logical reasoning to train object detection models with arbitrary types of weak supervision. We empirically show on different datasets that using all available information is beneficial as our ProbKT leads to significant improvement on target domain and better generalisation compared to existing baselines. We also showcase the ability of our approach to handle complex logic statements as supervision signal.

****

A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification

Paul F Jaeger,Carsten Tim Lüth,Lukas Klein,Till J. Bungert

Reliable application of machine learning-based decision systems in the wild is one of the major challenges currently investigated by the field. A large portion of established approaches aims to detect erroneous predictions by means of assigning confidence scores. This confidence may be obtained by either quantifying the model's predictive uncertainty, learning explicit scoring functions, or assessing whether the input is in line with the training distribution. Curiously, while these approaches all state to address the same eventual goal of detecting failures of a classifier upon real-world application, they currently constitute largely separated research fields with individual evaluation protocols, which either exclude a substantial part of relevant methods or ignore large parts of relevant failure sources. In this work, we systematically reveal current pitfalls caused by these inconsistencies and derive requirements for a holistic and realistic evaluation of failure detection. To demonstrate the relevance of this unified perspective, we present a large-scale empirical study for the first time enabling benchmarking confidence scoring functions w.r.t all relevant methods and failure sources. The revelation of a simple softmax response baseline as the overall best performing method underlines the drastic shortcomings of current evaluation in the plethora of publicized research on confidence scoring. Code and trained models are at https://github.com/https://github.com/IML-DKFZ/fd-shifts

****

Signs in the Lottery: Structural Similarities Between Winning Tickets

Isabel Holler,Mats Leon Richter,Ulf Krumnack

Winning tickets are sparse subnetworks of a deep network that can be trained in isolation to the same performance as the full network. Winning tickets have been found in many different contexts, however their structural characteristics are not well understood. We propose that the signs of the connections in winning tickets play a crucial role. We back this claim by introducing a sign-based structural comparison
metric that allows to distinguish winning tickets from other sparse networks. We further analyze typical (signed) patterns in convolutional kernels of winning tickets and find structures that resemble patterns found in trained networks.

****

Computational Doob h-transforms for Online Filtering of Discretely Observed Diffusions

Nicolas Chopin,Andras Fulop,Jeremy Heng,Alexandre H. Thiery

This paper is concerned with online filtering of discretely observed nonlinear diffusion processes. Our approach is based on the fully adapted auxiliary particle filter, which involves Doob's h-transforms that are typically intractable. We propose a computational framework to approximate these h-transforms by solving the underlying backward Kolmogorov equations using nonlinear Feynman-Kac formulas and neural networks. The methodology allows one to train a locally optimal particle filter prior to the data-assimilation procedure. Numerical experiments illustrate that the proposed approach can be orders of magnitude more efficient than the bootstrap particle filter in the regime of highly informative observations, when the observations are extreme under the model, and if the state dimension is large.

****

Reconciling feature sharing and multiple predictions with   MIMO Vision Transformers

Rémy Sun,Clément Masson,Nicolas THOME,Matthieu Cord

Multi-input multi-output training improves network performance by optimizing multiple subnetworks simultaneously. In this paper, we propose MixViT, the first MIMO framework for vision transformers that takes advantage of ViTs' innate mechanisms to share features between subnetworks. This is in stark contrast to traditional MIMO CNNs that are limited by their inability to mutualize features. Unlike them, MixViT only separates subnetworks in the last layers thanks to a novel source attribution that ties tokens to specific subnetworks. As such, we retain the benefits of multi-output supervision while training strong features useful to both subnetworks. We verify MixViT leads to significant gains across multiple architectures (ConViT, CaiT) and datasets (CIFAR, TinyImageNet, ImageNet-100, and ImageNet-1k) by fitting multiple subnetworks at the end of a base model.

****

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu,Arthur Gretton

We consider the estimation of average and counterfactual treatment effects, under two settings: back-door adjustment and front-door adjustment. The goal in both cases is to recover the treatment effect without having an access to a hidden confounder. This objective is attained by first estimating the conditional mean of the desired outcome variable given relevant covariates (the ``first stage" regression), and then taking the (conditional) expectation of this function as a ``second stage" procedure.

We propose to compute these conditional expectations directly using a regression function to the learned input features of the first stage, thus avoiding the need for sampling or density estimation. All functions and features (and in particular, the output features in the second stage) are neural networks learned adaptively from data, with the sole requirement that the final layer of the first stage should be linear. The proposed method is shown to converge to the true causal parameter, and outperforms the recent state-of-the-art methods on challenging causal benchmarks, including settings involving high-dimensional image data.

****

Chopping Formers is what you need in Vision

Francesca Babiloni,Thomas Tanay,Matteo Maggioni,Jiankang Deng,Ales Leonardis,Stefanos Zafeiriou

This work presents a new dynamic and fully-connected layer (DFC) that generalizes existing layers and is free from hard inductive biases. Then, it describes how to factorize the DFC weights efficiently.

Using the Einstein convention as framework, we define the DFC as a fully connected layer with the weight tensor created as a function of the input. DFC is the non-linear extension of the most general case of linear layer for neural network, and therefore all major neural network layers, from convolution to self-attention, are particular cases of DFCs. A stack of DFCs interleaved by non-linearities defines a new super-class of neural networks: \emph{Formers}.

DFC has four major characteristics: it is Dynamic and Spatially Adaptive, it has a Global Receptive Field, and it mixes all the available channels' information.

In their complete form, DFCs are powerful layers free from hard inductive biases, but their use is limited in practice by their prohibitive computational cost. To overcome this limitation and deploy DFC in real computer-vision applications, we propose to use the CP decomposition, showing that it is possible to factorize the DFC layer into smaller, manageable blocks without losing any representational power. Finally, we propose ChoP'D Former, an architecture making use of a new decomposition of the DFC layer into five sequential operations, each incorporating one characteristic of the original DFC tensor. Chop'D Former leverages dynamic gating and integral image, achieves global spatial reasoning with constant time complexity, and has a receptive field that can adapt depending on the task. Extensive experiments demonstrate that our ChoP'D Former is competitive with state-of-the-art results on three well-known computer vision benchmarks, namely Large-Scale Classification, Object Detection, and Instance Segmentation, suppressing the need for expensive architecture search and hyperparameter optimization.

****

Towards Estimating Transferability using Hard Subsets

Tarun Ram Menta,Surgan Jandial,Akash Patil,Vimal K B,Saketh Bachu,Balaji Krishnamurthy,Vineeth N. Balasubramanian,Chirag Agarwal,Mausoom Sarkar

As transfer learning techniques are increasingly used to transfer knowledge from the source model to the target task, it becomes important to quantify which source models are suitable for a given target task without performing computationally expensive fine-tuning. In this work, we propose HASTE (HArd Subset TransfErability), a new strategy to estimate the transferability of a source model to a particular target task using only a harder subset of target data. By leveraging the model’s internal and output representations, we introduce two techniques – one class-agnostic and another class-specific – to identify harder subsets and show that HASTE can be used with any existing transferability metric to improve their reliability. We further analyze the relation between HASTE and the optimal average log-likelihood as well as negative conditional entropy and empirically validate our theoretical bounds. Our experimental results across multiple source model architectures, target datasets, and transfer learning tasks show that HASTE-modified metrics are consistently better or on par with the state-of-the-art transferability metrics.

****

Data Pricing Mechanism Based on Property Rights Compensation Distribution

shiqian Liu,Peizheng Wang,Chao Wu

While machine learning (ML) benefits from data, it also faces the challenges of ambiguous data ownership, including privacy violations and increased costs of using data. Yet existing approaches to data valuation may focus on preventing privacy breaches, but do not truly protect data ownership. This is because a data trading marketplace that protects data ownership should achieve this goal: once data is traded, its ownership does not transfer to a new owner but merely enlarges its coverage. Considering that the transfer of property rights in the process of data trading makes compensation necessary, this paper proposes the first data valuation mechanism based on modern property rights theory. Specifically, we propose the integration of property rights to improve the final revenue of the entire workflow called the “data chain” while compensating process executors who lost ownership after integration. Then, we consider the expectations of both the integrator and the integrated party during the compensation allocation. For the former, we apply compound interest to assess a total compensation equivalent to the time value for the Data chain. For the latter, we respect and meet their expectations as much as possible. To achieve this, we provide the framework based on Least-core to assign the compensation and prove that our framework can also work compared to existing algorithms. Finally, to cope with more complex situations, we adjust the traditional Least-core and demonstrate theoretically and experimentally that the compensation mechanism is feasible and effective in solving the data pricing problem.

****

Knowledge-Driven Active Learning

Gabriele Ciravegna,Frederic Precioso,Marco Gori

The deployment of Deep Learning (DL) models is still precluded in those contexts where the amount of supervised data is limited. To answer this issue, active learning strategies aim at minimizing the amount of labelled data required to train a DL model. Most active strategies are based on uncertain sample selection, and even often restricted to samples lying close to the decision boundary. These techniques are theoretically sound, but an understanding of the selected samples based on their content is not straightforward, further driving non-experts to consider DL as a black-box. For the first time, here we propose a different approach, taking into consideration common domain-knowledge and enabling non-expert users to train a model with fewer samples. In our Knowledge-driven Active Learning (KAL) framework, rule-based knowledge is converted into logic constraints and their violation is checked as a natural guide for sample selection. We show that even simple relationships among data and output classes offer a way to spot predictions for which the model need supervision. The proposed approach (i) outperforms many active learning strategies in terms of average F1 score, particularly in those contexts where domain knowledge is rich. Furthermore, we empirically demonstrate that (ii) KAL discovers data distribution lying far from the initial training data unlike uncertainty-based strategies, (iii) it ensures domain experts that the provided knowledge is respected by the model on test data, and (iv) it can be employed even when domain-knowledge is not available by coupling it with a XAI technique. Finally, we also show that KAL is also suitable for object recognition tasks and, its computational demand is low, unlike many recent active learning strategies.

****

TranSpeech: Speech-to-Speech Translation With Bilateral Perturbation

Rongjie Huang,Jinglin Liu,Huadai Liu,Yi Ren,Lichao Zhang,Jinzheng He,Zhou Zhao

Direct speech-to-speech translation (S2ST) with discrete units leverages recent progress in speech representation learning. Specifically, a sequence of discrete representations derived in a self-supervised manner are predicted from the model and passed to a vocoder for speech reconstruction, while still facing the following challenges: 1) Acoustic multimodality: the discrete units derived from speech with same content could be indeterministic due to the acoustic property (e.g., rhythm, pitch, and energy), which causes deterioration of translation accuracy; 2) high latency: current S2ST systems utilize autoregressive models which predict each unit conditioned on the sequence previously generated, failing to take full advantage of parallelism. In this work, we propose TranSpeech, a speech-to-speech translation model with bilateral perturbation. To alleviate the acoustic multimodal problem, we propose bilateral perturbation (BiP), which consists of the style normalization and information enhancement stages, to learn only the linguistic information from speech samples and generate more deterministic representations. With reduced multimodality, we step forward and become the first to establish a non-autoregressive S2ST technique, which repeatedly masks and predicts unit choices and produces high-accuracy results in just a few cycles. Experimental results on three language pairs demonstrate that BiP yields an improvement o f 2.9 BLEU on average compared with a baseline textless S2ST model. Moreover, our parallel decoding shows a significant reduction of inference latency, enabling speedup up to 21.4x than autoregressive technique. Audio samples are available at https://TranSpeech.github.io

****

D4FT: A Deep Learning Approach to Kohn-Sham Density Functional Theory

Tianbo Li,Min Lin,Zheyuan Hu,Kunhao Zheng,Giovanni Vignale,Kenji Kawaguchi,A.H. Castro Neto,Kostya S. Novoselov,Shuicheng YAN

Kohn-Sham Density Functional Theory (KS-DFT) has been traditionally solved by the Self-Consistent Field (SCF) method. Behind the SCF loop is the physics intuition of solving a system of non-interactive single-electron wave functions under an effective potential. In this work, we propose a deep learning approach to KS-DFT. First, in contrast to the conventional SCF loop, we propose to directly minimize the total energy by reparameterizing the orthogonal constraint as a feed-forward computation. We prove that such an approach has the same expressivity as the SCF method, yet reduces the computational complexity from O(N^4) to O(N^3). Second, the numerical integration which involves a summation over the quadrature grids can be amortized to the optimization steps. At each step, stochastic gradient descent (SGD) is performed with a sampled minibatch of the grids. Extensive experiments are carried out to demonstrate the advantage of our approach in terms of efficiency and stability. In addition, we show that our approach enables us to explore more complex neural-based wave functions.

****

Warping the Space: Weight Space Rotation for Class-Incremental Few-Shot Learning

Do-Yeon Kim,Dong-Jun Han,Jun Seo,Jaekyun Moon

Class-incremental few-shot learning, where new sets of classes are provided sequentially with only a few training samples, presents a great challenge due to catastrophic forgetting of old knowledge and overfitting caused by lack of data. During finetuning on new classes, the performance on previous classes deteriorates quickly even when only a small fraction of parameters are updated, since the previous knowledge is broadly associated with most of the model parameters in the original parameter space. In this paper, we introduce WaRP, the \textit{weight space rotation process}, which transforms the original parameter space into a new space so that we can push most of the previous knowledge compactly into only a few important parameters. By properly identifying and freezing these key parameters in the new weight space, we can finetune the remaining parameters without affecting the knowledge of previous classes. As a result, WaRP provides an additional room for the model to effectively learn new classes in future incremental sessions. Experimental results confirm the effectiveness of our solution and show the improved performance over the state-of-the-art methods.

****

Learning to Reason and Act in Cascading Processes

Yuval Atzmon,Eli Meirom,Shie Mannor,Gal Chechik

Training agents to control a dynamic environment is a fundamental task in AI. In many environments the dynamics can be summarized by a small set of events that capture the semantic behavior of the system. Typically, these events form chains or cascades. We often wish to change the system behavior using a single intervention that propagates through the cascade. For instance, one may trigger a biochemical cascade to switch the state of a cell, or reroute a truck in logistic chains to meet an unexpected, urgent delivery.

We introduce a new supervised learning setup called "Cascade". An agent observes a system with a known dynamics evolving from some initial state. It is given a structured semantic instruction and needs to make an intervention that triggers a cascade of events, such that the system reaches an alternative (counterfactual) behavior. We provide a test-bed for this problem, consisting of physical objects.

We combine semantic tree search with an event-driven forward model and devise an algorithm that learns to efficiently search in exponentially large semantic trees of continuous spaces. We demonstrate that our approach learns to effectively follow instructions to intervene in previously unseen complex scenes. When provi ded an observed cascade of events, it can also reason about alternative outcomes.

****

Over-parameterized Model Optimization with Polyak-{\L}ojasiewicz Condition
Yixuan Chen,Yubin Shi,Mingzhi Dong,Xiaochen Yang,Dongsheng Li,Yujiang Wang,Robert Dick,Qin Lv,Yingying Zhao,Fan Yang,Ning Gu,Li Shang
This work pursues the optimization of over-parameterized deep models for superior training efficiency and test performance. We first theoretically emphasize the importance of two properties of over-parameterized models, i.e., the convergence gap and the generalization gap. Subsequent analyses unveil that these two gaps can be upper-bounded by the ratio of the Lipschitz constant and the Polyak-{\L}ojasiewicz (PL) constant, a crucial term abbreviated as the \emph{condition number}. Such discoveries have led to a structured pruning method with a novel pruning criterion. That is, we devise a gating network that dynamically detects and masks out those poorly-behaved nodes of a deep model during the training session. To this end, this gating network is learned via minimizing the \emph{condition number} of the target model, and this process can be implemented as an extra regularization loss term. Experimental studies demonstrate that the proposed method outperforms the baselines in terms of both training efficiency and test performance, exhibiting the potential of generalizing to a variety of deep network architectures and tasks.

****

Differentially private Bias-Term Only Fine-tuning of Foundation Models
Zhiqi Bu,Yu-Xiang Wang,Sheng Zha,George Karypis
We study the problem of differentially private (DP) fine-tuning of large pre-trained models — a recent privacy-preserving approach suitable for solving downstream tasks with sensitive data. Existing work has demonstrated that high accuracy is possible under strong privacy constraint, yet requires significant computational overhead or modifications to the network architecture.

We propose differentially private bias-term fine-tuning (DP-BiTFiT), which matches the state-of-the-art accuracy for DP algorithms and the efficiency of the standard BiTFiT. DP-BiTFiT is model agnostic (not modifying the network architecture), parameter efficient (only training about 0.1% of the parameters), and computation efficient (almost removing the overhead caused by DP, in both the time and space complexity). On a wide range of tasks, DP-BiTFiT is $2\sim 30\times$ faster and uses $2\sim 8\times$ less memory than DP full fine-tuning, even faster than the standard full fine-tuning. This amazing efficiency enables us to conduct DP fine-tuning on language and vision tasks with long-sequence texts and high-resolution images, which were computationally difficult using existing methods.

****

Jointly Learning Visual and Auditory Speech Representations from Raw Data
Alexandros Haliassos,Pingchuan Ma,Rodrigo Mira,Stavros Petridis,Maja Pantic
We present RAVEn, a self-supervised multi-modal approach to jointly learn visual and auditory speech representations. Our pre-training objective involves encoding masked inputs, and then predicting contextualised targets generated by slowly-evolving momentum encoders. Driven by the inherent differences between video and audio, our design is asymmetric w.r.t. the two modalities' pretext tasks: Whereas the auditory stream predicts both the visual and auditory targets, the visual one predicts only the auditory targets. We observe strong results in low- and high-resource labelled data settings when fine-tuning the visual and auditory encoders resulting from a single pre-training stage, in which the encoders are jointly trained. Notably, RAVEn surpasses all self-supervised methods on visual speech recognition (VSR) on LRS3, and combining RAVEn with self-training using only 30 hours of labelled data even outperforms a recent semi-supervised method trained on 90,000 hours of non-public data. At the same time, we achieve state-of-the-art results in the LRS3 low-resource setting for auditory speech recognition (as well as for VSR). Our findings point to the viability of learning powerful speech representations entirely from raw video and audio, i.e., without relying on handcrafted features. Code and models are available at https://github.com/ahali

assos/raven.
****

Diminishing Return of Value Expansion Methods in Model-Based Reinforcement Learning

Daniel Palenicek,Michael Lutter,Joao Carvalho,Jan Peters

Model-based reinforcement learning is one approach to increase sample efficiency. However, the accuracy of the dynamics model and the resulting compounding error over modelled trajectories are commonly regarded as key limitations. A natural question to ask is: How much more sample efficiency can be gained by improving the learned dynamics models? Our paper empirically answers this question for the class of model-based value expansion methods in continuous control problems. Value expansion methods should benefit from increased model accuracy by enabling longer rollout horizons and better value function approximations. Our empirical study, which leverages oracle dynamics models to avoid compounding model errors, shows that (1) longer horizons increase sample efficiency, but the gain in improvement decreases with each additional expansion step, and (2) the increased model accuracy only marginally increases the sample efficiency compared to learned models with identical horizons. Therefore, longer horizons and increased model accuracy yield diminishing returns in terms of sample efficiency. These improvements in sample efficiency are particularly disappointing when compared to model-free value expansion methods. Even though they introduce no computational overhead, we find their performance to be on-par with model-based value expansion methods. Therefore, we conclude that the limitation of model-based value expansion methods is not the model accuracy of the learned models. While higher model accuracy is beneficial, our experiments show that even a perfect model will not provide an un-rivaled sample efficiency but that the bottleneck lies elsewhere.
****

Differentially Private Optimization on Large Model at Small Cost

Zhiqi Bu,Yu-Xiang Wang,Sheng Zha,George Karypis

Differentially private (DP) optimization is the standard paradigm to learn large neural networks that are accurate and privacy-preserving. The computational cost for DP deep learning, however, is notoriously heavy due to the per-sample gradient clipping. Existing DP implementations are $2-1000\times$ more costly in time and space complexity than the standard (non-private) training. In this work, we develop a novel Book-Keeping (BK) technique that implements existing DP optimizers (thus achieving the same accuracy), with a substantial improvement on the computational cost. Specifically, BK enables DP training on large models and high dimensional data to be roughly as efficient as the standard training, whereas previous DP algorithms can be inefficient or incapable of training due to memory error. The computational advantage of BK is supported by the complexity analysis as well as extensive experiments on vision and language tasks. Our implementation achieves state-of-the-art (SOTA) accuracy with very small extra cost: on GPT2 and at the same memory cost, BK has 1.0$\times$ the time complexity of the standard training (0.75$\times$ training speed in practice), and 0.6$\times$ the time complexity of the most efficient DP implementation (1.24$\times$ training speed in practice). We will open-source the codebase for the BK algorithm.
****

CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment

Hongwei Xue,Yuchong Sun,Bei Liu,Jianlong Fu,Ruihua Song,Houqiang Li,Jiebo Luo

Pre-trained image-text models, like CLIP, have demonstrated the strong power of vision-language representation learned from a large scale of web-collected image-text data. In light of the well-learned visual features, there are works that transfer image representation to the video domain and achieve good results. However, adapting image-text pre-trained models to video-text pre-training (i.e., post-pretraining) has not demonstrated a significant advantage yet. In this paper, we tackle this challenge by raising and addressing two questions: 1) what are the factors hindering post-pretraining CLIP from improving performance on video-text tasks, and 2) how to mitigate the impact of these factors. Through a series of comparative experiments and analyses, we find that the data scale and domain gap between language sources have large impacts. By these observations, we propos

e an Omnisource Cross-modal Learning method equipped with a Video Proxy mechanism on the basis of CLIP, namely CLIP-ViP. Extensive results show that our approach improves the performance of CLIP on video-text retrieval by a large margin. Our model achieves state-of-the-art results on a variety of datasets, including MSR-VTT, DiDeMo, LSMDC, and ActivityNet. We release our code and pre-trained CLIP-ViP models at \url{https://github.com/microsoft/XPretrain/tree/main/CLIP-ViP}.
****
Pre-training via Denoising for Molecular Property Prediction
Sheheryar Zaidi,Michael Schaarschmidt,James Martens,Hyunjik Kim,Yee Whye Teh,Alvaro Sanchez-Gonzalez,Peter Battaglia,Razvan Pascanu,Jonathan Godwin
Many important problems involving molecular property prediction from 3D structures have limited data, posing a generalization challenge for neural networks. In this paper, we describe a pre-training technique based on denoising that achieves a new state-of-the-art in molecular property prediction by utilizing large datasets of 3D molecular structures at equilibrium to learn meaningful representations for downstream tasks. Relying on the well-known link between denoising autoencoders and score-matching, we show that the denoising objective corresponds to learning a molecular force field -- arising from approximating the Boltzmann distribution with a mixture of Gaussians -- directly from equilibrium structures. Our experiments demonstrate that using this pre-training objective significantly improves performance on multiple benchmarks, achieving a new state-of-the-art on the majority of targets in the widely used QM9 dataset. Our analysis then provides practical insights into the effects of different factors -- dataset sizes, model size and architecture, and the choice of upstream and downstream datasets -- on pre-training.
****
Equivariant Energy-Guided SDE for Inverse Molecular Design
Fan Bao,Min Zhao,Zhongkai Hao,Peiyao Li,Chongxuan Li,Jun Zhu
Inverse molecular design is critical in material science and drug discovery, where the generated molecules should satisfy certain desirable properties. In this paper, we propose equivariant energy-guided stochastic differential equations (EEGSDE), a flexible framework for controllable 3D molecule generation under the guidance of an energy function in diffusion models. Formally, we show that EEGSDE naturally exploits the geometric symmetry in 3D molecular conformation, as long as the energy function is invariant to orthogonal transformations. Empirically, under the guidance of designed energy functions, EEGSDE significantly improves the baseline on QM9, in inverse molecular design targeted to quantum properties and molecular structures. Furthermore, EEGSDE is able to generate molecules with multiple target properties by combining the corresponding energy functions linearly.
****
Contrastive Value Learning: Implicit Models for Simple Offline RL
Bogdan Mazoure,Benjamin Eysenbach,Ofir Nachum,Jonathan Tompson
Model-based reinforcement learning (RL) methods are appealing in the offline setting because they allow an agent to reason about the consequences of actions without interacting with the environment. Prior methods learn a 1-step dynamics model, which predicts the next state given the current state and action. These models do not immediately tell the agent which actions to take, but must be integrated into a larger RL framework. Can we model the environment dynamics in a different way, such that the learned model does directly indicate the value of each action? In this paper, we propose Contrastive Value Learning (CVL), which learns an implicit, multi-step model of the environment dynamics. This model can be learned without access to reward functions, but nonetheless can be used to directly estimate the value of each action, without requiring any TD learning. Because this model represents the multi-step transitions implicitly, it avoids having to predict high-dimensional observations and thus scales to high-dimensional tasks. Our experiments demonstrate that CVL outperforms prior offline RL methods on complex continuous control benchmarks.
****
Vectorial Graph Convolutional Networks

ZhongYu Li,Geng Zhao,Hao Ning

Graph Convolutional Networks (GCN) have drawn considerable attention recently due to their outstanding performance in processing graph-structured data. However, GCNs still limited to the undirected graph because they theoretically require a symmetric matrix as the basis for the Laplacian transform. This causes the isotropic problem of the operator and reduced sensitivity in response to different information. In order to solve the problem, we generalize the spectral convolution operator to directed graphs by field extension, which improves the edge representations from scalars to vectors. Therefore, it brings in the concept of direction. That is to say, and even homogeneous information can become distinguishable by its differences in directions.In this paper, we propose the Vectorial Graph Convolutional Network(VecGCN) and the experimental evidence showing the advantages of a variety of directed graph node classification and link prediction tasks.

****

Traversing Between Modes in Function Space for Fast Ensembling

Eunggu Yun,Hyungi Lee,Giung Nam,Juho Lee

Deep ensemble is a simple yet powerful way to improve the performance of deep neural networks. Under this motivation, recent works on mode connectivity have shown that parameters of ensembles are connected by low-loss subspaces, and one can efficiently collect ensemble parameters in those subspaces. While this provides a way to efficiently train ensembles, for inference, one should still execute multiple forward passes using all the ensemble parameters, which often becomes a serious bottleneck for real-world deployment. In this work, we propose a novel framework to reduce such costs. Given a low-loss subspace connecting two modes of a neural network, we build an additional neural network predicting outputs of the original neural network evaluated at a certain point in the low-loss subspace. The additional neural network, what we call a “ bridge”, is a lightweight network taking minimal features from the original network, and predicting outputs for the low-loss subspace without forward passes through the original network. We empirically demonstrate that we can indeed train such bridge networks and significantly reduce inference costs with the help of the bridge networks.

****

Poisson Process for Bayesian Optimization

Xiaoxing Wang,Jiaxing Li,Chao Xue,Wei Liu,Chaoyue Wang,Weifeng Liu,Xiaokang Yang,Junchi Yan,Dacheng Tao

Bayesian Optimization (BO) is a sample-efficient, model-based method for optimizing black-box functions which can be expensive to evaluate. Traditionally, BO seeks a probabilistic surrogate model, such as Tree-structured Parzen Estimator (TPE), Sequential Model Algorithm Configuration (SMAC), and Gaussian process (GP), based on the exact observed values. However, compared to the value response, relative ranking is hard to be disrupted due to noise resulting in better robustness. Moreover, it has better practicality when the exact value responses are intractable, but information about candidate preferences can be acquired. Thus, this work introduces an efficient BO framework, named PoPBO, consisting of a novel ranking-based response surface based on Poisson process and two acquisition functions to accommodate the proposed surrogate model. We show empirically that PoPBO improves efficacy and efficiency on both simulated and real-world benchmarks, including HPO and NAS.

****

DPMAC: Differentially Private Communication for Cooperative Multi-Agent Reinforcement Learning

Canzhe Zhao,Yanjie Ze,Jing Dong,Baoxiang Wang,Shuai Li

Communication lays the foundation for cooperation in human society and in multi-agent reinforcement learning (MARL). Humans also desire to maintain their privacy when communicating with others, yet such privacy concern has not been considered in existing works in MARL. We propose the \textit{differentially private multi-agent communication} (DPMAC) algorithm, which protects the sensitive information of individual agents by equipping each agent with a local message sender with rigorous (ϵ, δ)-differential privacy (DP) guarantee. In contrast to directly perturbing the messages with predefined DP noise as commonly done in privacy-preserving scenarios, we adopt a stochastic message sender for each agent respectively and incorporate the DP requirement into the sender, which automatically adjusts the learned message distribution to alleviate the instability caused by DP noise. Further, we prove the existence of a Nash equilibrium in cooperative MARL with privacy-preserving communication, which suggests that this problem is game-theoretically learnable. Extensive experiments demonstrate a clear advantage of DPMAC over baseline methods in privacy-preserving scenarios.

****

Q-learning with regularization converges with non-linear non-stationary features

Diogo S. Carvalho,Francisco S. Melo,Pedro A. Santos

The deep Q-learning architecture is a neural network composed of non-linear hidden layers that learn features of states and actions and a final linear layer that learns the Q-values of the features. The parameters of both components can possibly diverge. Regularization of the updates is known to solve the divergence problem of fully linear architectures, where features are stationary and known a priori. We propose a deep Q-learning scheme that uses regularization of the final linear layer of architecture, updating it along a faster time-scale, and stochastic full-gradient descent updates for the non-linear features at a slower time-scale. We prove the proposed scheme converges with probability 1. Finally, we provide a bound on the error introduced by regularization of the final linear layer of the architecture.

****

On the Feasibility of Cross-Task Transfer with Model-Based Reinforcement Learning

Yifan Xu,Nicklas Hansen,Zirui Wang,Yung-Chieh Chan,Hao Su,Zhuowen Tu

Reinforcement Learning (RL) algorithms can solve challenging control problems directly from image observations, but they often require millions of environment interactions to do so. Recently, model-based RL algorithms have greatly improved sample-efficiency by concurrently learning an internal model of the world, and supplementing real environment interactions with imagined rollouts for policy improvement. However, learning an effective model of the world from scratch is challenging, and in stark contrast to humans that rely heavily on world understanding and visual cues for learning new skills. In this work, we investigate whether internal models learned by modern model-based RL algorithms can be leveraged to solve new, distinctly different tasks faster. We propose Model-Based Cross-Task Transfer (XTRA), a framework for sample-efficient online RL with scalable pretraining and finetuning of learned world models. By offline multi-task pretraining and online cross-task finetuning, we achieve substantial improvements over a baseline trained from scratch; we improve mean performance of model-based algorithm EfficientZero by 23%, and by as much as 71% in some instances. Project page: https://nicklashansen.github.io/xtra

****

Fast and Precise: Adjusting Planning Horizon with Adaptive Subgoal Search

Micha■ Zawalski,Micha■ Tyrolski,Konrad Czechowski,Tomasz Odrzygó■d■,Damian Stachura,Piotr Pi■kos,Yuhuai Wu,■ukasz Kuci■ski,Piotr Mi■o■

Complex reasoning problems contain states that vary in the computational cost required to determine the right action plan. To take advantage of this property, we propose Adaptive Subgoal Search (AdaSubS), a search method that adaptively adjusts the planning horizon. To this end, AdaSubS generates diverse sets of subgoals at different distances. A verification mechanism is employed to filter out unreachable subgoals swiftly, making it possible to focus on feasible further subgoals. In this way, AdaSubS benefits from the efficiency of planning with longer-term subgoals and the fine control with shorter-term ones, and thus scales well to difficult planning problems. We show that AdaSubS significantly surpasses hierarchical planning algorithms on three complex reasoning tasks: Sokoban, the Rubik’s Cube, and the inequality-proving benchmark INT.

****

A Simple Yet Powerful Deep Active Learning With Snapshots Ensembles

Seohyeon Jung,Sanghyun Kim,Juho Lee

Given an unlabeled pool of data and the experts who can label them, active learning aims to build an agent that can effectively acquire data to be queried to the experts, maximizing the gain in performance when trained with them. While there are several principles for active learning, a prevailing approach is to estimate uncertainties of predictions for unlabeled samples and use them to define acquisition functions. Active learning with the uncertainty principle works well for deep learning, especially for large-scale image classification tasks with deep neural networks. Still, it is often overlooked how the uncertainty of predictions is estimated, despite the common findings on the difficulty of accurately estimating uncertainties of deep neural networks. In this paper, we highlight the effectiveness of snapshot ensembles for deep active learning. Compared to the previous approaches based on Monte-Carlo dropout or deep ensembles, we show that a simple acquisition strategy based on uncertainties estimated from parameter snapshots gathered from a single optimization path significantly improves the quality of the acquired samples. Based on this observation, we further propose an efficient active learning algorithm that maintains a single learning trajectory throughout the entire active learning episodes, unlike the existing algorithms training models from scratch for every active learning episode. Through the extensive empirical comparison, we demonstrate the effectiveness of snapshot ensembles for deep active learning.

****

Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk,Dennis Frauen,Stefan Feuerriegel

Existing machine learning methods for causal inference usually estimate quantities expressed via the mean of potential outcomes (e.g., average treatment effect). However, such quantities do not capture the full information about the distribution of potential outcomes. In this work, we estimate the density of potential outcomes after interventions from observational data. For this, we propose a novel, fully-parametric deep learning method called Interventional Normalizing Flows. Specifically, we combine two normalizing flows, namely (i) a teacher flow for estimating nuisance parameters and (ii) a student flow for a parametric estimation of the density of potential outcomes. We further develop a tractable optimization objective via a one-step bias correction for an efficient and doubly robust estimation of the student flow parameters. As a result our Interventional Normalizing Flows offer a properly normalized density estimator. Across various experiments, we demonstrate that our Interventional Normalizing Flows are expressive and highly effective, and scale well with both sample size and high-dimensional confounding. To the best of our knowledge, our Interventional Normalizing Flows are the first fully-parametric, deep learning method for density estimation of potential outcomes.

****

Backdoor or Feature? A New Perspective on Data Poisoning

Alaa Khaddaj,Guillaume Leclerc,Aleksandar Makelov,Kristian Georgiev,Andrew Ilyas,Hadi Salman,Aleksander Madry

In a backdoor attack, an adversary adds maliciously constructed ("backdoor") examples into a training set to make the resulting model vulnerable to manipulation. Defending against such attacks---that is, finding and removing the backdoor examples---typically involves viewing these examples as outliers and using techniques from robust statistics to detect and remove them.

In this work, we present a new perspective on backdoor attacks. We argue that without structural information on the training data distribution, backdoor attacks are indistinguishable from naturally-occuring features in the data (and thus impossible to ``detect'' in a general sense). To circumvent this impossibility, we assume that a backdoor attack corresponds to the strongest feature in the training data. Under this assumption---which we make formal---we develop a new framework for detecting backdoor attacks. Our framework naturally gives rise to a corresponding algorithm whose efficacy we show both theoretically and experimentally.

****

A Curriculum Perspective to Robust Loss Functions

Zebin Ou,Yue Zhang

Learning with noisy labels is a fundamental problem in machine learning. Much work has been done in designing loss functions that are theoretically robust against label noise. However, it remains unclear why robust loss functions can underfit and why loss functions deviating from theoretical robustness conditions can appear robust. To elucidate these questions, we show that most robust loss functions differ only in the sample-weighting curriculums they implicitly define. The curriculum perspective enables straightforward analysis of the training dynamics with each loss function, which has not been considered in existing theoretical approaches. We show that underfitting can be attributed to marginal sample weights during training, and noise robustness can be attributed to larger weights for clean samples than noisy samples. With a simple fix to the curriculums, robust loss functions that severely underfit can become competitive with the state-of-the-art.

****

Decoupled Training for Long-Tailed Classification With Stochastic Representations

Giung Nam,Sunguk Jang,Juho Lee

Decoupling representation learning and classifier learning has been shown to be effective in classification with long-tailed data. There are two main ingredients in constructing a decoupled learning scheme; 1) how to train the feature extractor for representation learning so that it provides generalizable representations and 2) how to re-train the classifier that constructs proper decision boundaries by handling class imbalances in long-tailed data. In this work, we first apply Stochastic Weight Averaging (SWA), an optimization technique for improving the generalization of deep neural networks, to obtain better generalizing feature extractors for long-tailed classification. We then propose a novel classifier re-training algorithm based on stochastic representation obtained from the SWA-Gaussian, a Gaussian perturbed SWA, and a self-distillation strategy that can harness the diverse stochastic representations based on uncertainty estimates to build more robust classifiers. Extensive experiments on CIFAR10/100-LT, ImageNet-LT, and iNaturalist-2018 benchmarks show that our proposed method improves upon previous methods both in terms of prediction accuracy and uncertainty estimation.

****

IT-NAS: Integrating Lite-Transformer into NAS for Architecture Seletion

Zihao Sun,Yu Hu,Longxing Yang,Shun Lu,Jilin Mei,Yinhe Han

Neural Architecture Search (NAS) aims to search for the best network in the pre-defined search space. However, much work focuses on the search strategy but little on the architecture selection process. Despite the fact that the weight-sharing based NAS has promoted the search efficiency, we notice that the architecture selection is quite unstable or circuitous. For instance, the differentiable NAS may derive the suboptimal architecture due to the performance collapse caused by bi-level optimization, or the One-shot NAS requires sampling and evaluating a large number of candidate structures. Recently, the self-attention mechanism achieves better performance in terms of the long-range modeling capabilities. Considering that different operations are widely distributed in the search space, we suggest leveraging the self-attention mechanism to extract the relationship among them and to determine which operation is superior to others. Therefore, we integrate Lite-Transformer into NAS for architecture selection. Specifically, we regard the feature map of each candidate operation as distinct patches and feed them into the Lite-Transformer module along with an additional Indicator Token (called IT). The cross attention among various operations can be extracted by the self-attention mechanism, and the importance of each candidate operation is then shown by the softmax result between the query of indicator token (IT) and other values of operational tokens. We experimentally demonstrate that our framework can select the truly representative architecture in different search spaces and achieves 2.39% test error on CIFAR-10 in DARTS search space, and 24.1% test error on ImageNet in the ProxylessNAS search space, as well as the stable and better performance in NAS-Bench-201 search space and S1-S4 search spaces, outperforming state-of-the-art NAS methods.

****

Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC

Yilun Du,Conor Durkan,Robin Strudel,Joshua B. Tenenbaum,Sander Dieleman,Rob Fergus,Jascha Sohl-Dickstein,Arnaud Doucet,Will Sussman Grathwohl

Since their introduction, diffusion models have quickly become the prevailing approach to generative modeling in many domains. They can be interpreted as learning the gradients of a time-varying sequence of log-probability density functions. This interpretation has motivated classifier-based and classifier-free guidance as methods for post-hoc control of diffusion models. In this work, we build upon these ideas using the score-based interpretation of diffusion models, and explore alternative ways to condition, modify, and reuse diffusion models for tasks involving compositional generation and guidance. In particular, we investigate why certain types of composition fail using current techniques and present a number of solutions. We conclude that the sampler (not the model) is responsible for this failure and propose new samplers, inspired by MCMC, which enable successful compositional generation. Further, we propose an energy-based parameterization of diffusion models which enables the use of new compositional operators and more sophisticated, Metropolis-corrected samplers. Intriguingly we find these samplers lead to notable improvements in compositional generation across a wide variety of problems such as classifier-guided ImageNet modeling and compositional text-to-image generation.

****

Martingale Posterior Neural Processes

Hyungi Lee,Eunggu Yun,Giung Nam,Edwin Fong,Juho Lee

A Neural Process (NP) estimates a stochastic process implicitly defined with neural networks given a stream of data, rather than pre-specifying priors already known, such as Gaussian processes. An ideal NP would learn everything from data without any inductive biases, but in practice, we often restrict the class of stochastic processes for the ease of estimation. One such restriction is the use of a finite-dimensional latent variable accounting for the uncertainty in the functions drawn from NPs. Some recent works show that this can be improved with more “data-driven” source of uncertainty such as bootstrapping. In this work, we take a different approach based on the martingale posterior, a recently developed alternative to Bayesian inference. For the martingale posterior, instead of specifying prior-likelihood pairs, a predictive distribution for future data is specified. Under specific conditions on the predictive distribution, it can be shown that the uncertainty in the generated future data actually corresponds to the uncertainty of the implicitly defined Bayesian posteriors. Based on this result, instead of assuming any form of the latent variables, we equip a NP with a predictive distribution implicitly defined with neural networks and use the corresponding martingale posteriors as the source of uncertainty. The resulting model, which we name as Martingale Posterior Neural Process (MPNP), is demonstrated to outperform baselines on various tasks.

****

GuoFeng: A Discourse-aware Evaluation Benchmark for Language Understanding, Translation and Generation

Longyue Wang,Zefeng Du,DongHuai Liu,Deng Cai,Dian Yu,Haiyun Jiang,Yan Wang,Shuming Shi,Zhaopeng Tu

Modeling discourse -- the linguistic phenomena that go beyond individual sentences, is a fundamental and challenging problem in natural language processing (NLP). However, existing evaluation benchmarks mainly focus on the evaluation of inter-sentence properties and overlook important discourse phenomena that cross sentences. To bridge the gap, we propose a GuoFeng benchmark that can evaluate intra-sentence discourse properties across a diverse set of NLP tasks, covering understanding, translation, and generation. GuoFeng consists of 9 document-level testsets in the literature domain, which contain rich discourse phenomena (e.g. cohesion and coherence) in Chinese and/or English. For linguistic analysis, we also propose a diagnostic test suite that can examine whether the target models learn discourse knowledge. We evaluate 17 general- and in-domain models based on Transformer and advanced pre-training architectures, showing that fine-grained pretraining based on document-level training data consistently improves the modeling of discourse information. We will release the datasets, pretrained models, and leaderboard, which we hope can significantly facilitate research in this field.

****

Multi-View Independent Component Analysis with Shared and Individual Sources

Teodora Pandeva,Patrick Forré

Independent component analysis (ICA) is a blind source separation method for linear disentanglement of independent latent sources from observed data. We investigate the special setting of noisy linear ICA where the observations are split among different views, each receiving a mixture of shared and individual sources. We prove that the corresponding linear structure is identifiable, and the shared sources can be recovered, provided that sufficiently many diverse views and data points are available. To computationally estimate the sources, we optimize a constrained form of the joint log-likelihood of the observed data among all views. We show empirically that our objective recovers the sources in high dimensional settings, also in the case when the measurements are corrupted by noise. Finally, we apply the proposed model in a challenging real-life application, where the estimated shared sources from two large transcriptome datasets (observed data) provided by two different labs (two different views) lead to a more plausible representation of the underlying graph structure than existing baselines.

****

Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning

Pedro Pinto Santos,Diogo S. Carvalho,Miguel Serras Vasco,Francisco S. Melo,Alberto Sardinha,Pedro A. Santos,Ana Paiva

We introduce hybrid execution in multi-agent reinforcement learning (MARL), a new paradigm in which agents aim to successfully perform cooperative tasks with any communication level at execution time by taking advantage of information-sharing among the agents. Under hybrid execution, the communication level can range from a setting in which no communication is allowed between agents (fully decentralized), to a setting featuring full communication (fully centralized). To formalize our setting, we define a new class of multi-agent partially observable Markov decision processes (POMDPs) that we name hybrid-POMDPs, which explicitly models a communication process between the agents. We contribute MARO, an approach that combines an autoregressive predictive model to estimate missing agents' observations, and a dropout-based RL training scheme that simulates different communication levels during the centralized training phase. We evaluate MARO on standard scenarios and extensions of previous benchmarks tailored to emphasize the negative impact of partial observability in MARL. Experimental results show that our method consistently outperforms baselines, allowing agents to act with faulty communication while successfully exploiting shared information.

****

Towards Open Temporal Graph Neural Networks

Kaituo Feng,Changsheng Li,Xiaolu Zhang,JUN ZHOU

Graph neural networks (GNNs) for temporal graphs have recently attracted increasing attentions, where a common assumption is that the class set for nodes is closed. However, in real-world scenarios, it often faces the open set problem with the dynamically increased class set as the time passes by. This will bring two big challenges to the existing dynamic GNN methods: (i) How to dynamically propagate appropriate information in an open temporal graph, where new class nodes are often linked to old class nodes. This case will lead to a sharp contradiction. This is because typical GNNs are prone to make the embeddings of connected nodes become similar, while we expect the embeddings of these two interactive nodes to be distinguishable since they belong to different classes. (ii) How to avoid catastrophic knowledge forgetting over old classes when learning new classes occurred in temporal graphs. In this paper, we propose a general and principled learning approach for open temporal graphs, called OTGNet, with the goal of addressing the above two challenges. We assume the knowledge of a node can be disentangl ed into class-relevant and class-agnostic one, and thus explore a new message passing mechanism by extending the information bottleneck principle to only propagate class-agnostic knowledge between nodes of different classes, avoiding aggregating conflictive information. Moreover, we devise a strategy to select both important and diverse triad sub-graph structures for effective class-incremental learning. Extensive experiments on three real-world datasets of different domains demonstrate the superiority of our method, compared to the baselines.

****

Learning Discriminative Representations for Chromosome Classification with Small Datasets

Minghui Li,Haoxi Zhang,Renhao Zhou,Linfeng Yu

Chromosome classification is crucial for karyotype analysis in cytogenetics. Karyotype analysis is a fundamental approach for clinical cytogeneticists to identify numerical and structural chromosomal abnormalities. However, classifying chromosomes accurately and robustly in clinical application is still challenging due to: 1) rich deformations of chromosome shape, 2) similarity of chromosomes, and 3) imbalanced and insufficient labelled dataset. This paper proposes a novel pipeline for the automatic classification of chromosomes. Unlike existing methods, our approach is primarily based on learning meaningful data representations rather than only finding classification features in given samples. The proposed pipeline comprises three stages: The first stage extracts meaningful visual features of chromosomes by utilizing ResNet with triplet loss. The second stage optimizes features from stage one to obtain a linear discriminative representation via maximal coding rate reduction. It ensures the clusters representing different chromosome types are far away from each other while embeddings of the same type are close to each other in the cluster. The third stage is to identify chromosomes. Based on the meaningful feature representation learned in the previous stage, traditional machine learning algorithms such as SVM are adequate for the classification task. Evaluation results on a publicly available dataset show that our method achieves 97.22% accuracy and is better than state-of-the-art methods.

****

ViewCo: Discovering Text-Supervised Segmentation Masks via Multi-View Semantic Consistency

Pengzhen Ren,Changlin Li,Hang Xu,Yi Zhu,Guangrun Wang,Jianzhuang Liu,Xiaojun Chang,Xiaodan Liang

Recently, great success has been made in learning visual representations from text supervision, facilitating the emergence of text-supervised semantic segmentation. However, existing works focus on pixel grouping and cross-modal semantic alignment, while ignoring the correspondence among multiple augmented views of the same image. To overcome such limitation, we propose multi-View Consistent learning (ViewCo) for text-supervised semantic segmentation. Specifically, we first propose text-to-views consistency modeling to learn correspondence for multiple views of the same input image. Additionally, we propose cross-view segmentation consistency modeling to address the ambiguity issue of text supervision by contrasting the segment features of Siamese visual encoders. The text-to-views consistency benefits dense assignment of the visual features by encouraging different crops to align with the same text, while the cross-view segmentation consistency modeling provides additional self-supervision, overcoming the limitation of ambiguous text supervision for segmentation masks. Trained with large-scale image-text data, our model can directly segment objects of arbitrary categories in a zero-shot manner. Extensive experiments show that ViewCo outperforms state-of-the-art methods on average by up to 2.9%, 1.6%, and 2.4% mIoU on PASCAL VOC2012, PASCAL Context, and COCO, respectively.

****

Simplicity bias in 1-hidden layer neural networks

Depen Morwani,Praneeth Netrapalli,jatin batra,Karthikeyan Shanmugam,Prateek Jain

Recent works \citep{shah2020pitfalls,chen2021intriguing} have demonstrated that neural networks exhibit extreme \emph{simplicity bias} (SB). That is, they learn \emph{only the simplest} features to solve a task at hand, even in the presence of other, more robust but more complex features. Due to lack of a general and rigorous definition of \emph{features}, these works showcase SB on \emph{semi-synthetic} datasets such as Color-MNIST, MNIST-CIFAR where defining features is relatively easier.

In this work, we rigorously define as well as thoroughly establish SB for \emph{one hidden layer} neural networks. More concretely, (i) we define SB as the network essentially being a function of a low dimensional projection of the inputs (ii) theoretically, we show that when the data is linearly separable, the network primarily depends on only the linearly separable (1-dimensional) subspace even in the presence of an arbitrarily large number of other, more complex features which could have led to a significantly more robust classifier, (iii) empirically, we show that models trained on \emph{real} datasets such as Imagenette and Waterbirds-Landbirds indeed depend on a low dimensional projection of the inputs, thereby demonstrating SB on these datasets, iv) finally, we present a natural ensemble approach that encourages diversity in models by training successive models on features not used by earlier models, and demonstrate that it yields models that are significantly more robust to Gaussian noise.
****

FedEED: Efficient Federated Distillation with Ensemble of Aggregated Models
Ho Man Kwan,Shenghui Song
In this paper, we study the key components of the knowledge distillation-based model aggregation in federated learning (FL). We first propose a generalized distillation framework where the process of federated distillation is divided into three key stages. By investigating the contributions of each stage, we introduce a new FL framework, named Federated Efficient Ensemble Distillation (FedEED), where the ensemble teacher is created based on aggregated models. Experiment results showed that FedEED outperforms the state-of-the-art methods, including FedAvg and FedDF, on the benchmark datasets. Besides performance, FedEED also demonstrated improved scalability and privacy when compared with existing distillation-based aggregation algorithms. In particular, FedEED does not require direct access to users' model, which can protect the users' privacy. Furthermore, due to the ensemble created by aggregated models, FedEED is highly scalable, and the asymmetric distillation scheme allows parallelism between server-side distillation and clients-side local training, which could speed up the training of large scale learning system.
****

Critical Batch Size Minimizes Stochastic First-Order Oracle Complexity of Deep Learning Optimizer using Hyperparameters Close to One
Hideaki Iiduka
Practical results have shown that deep learning optimizers using small constant learning rates, hyperparameters close to one, and large batch sizes can find the model parameters of deep neural networks that minimize the loss functions. We first show theoretical evidence that the momentum method (Momentum) and adaptive moment estimation (Adam) perform well in the sense that the upper bound of the theoretical performance measure is small with a small constant learning rate, hyperparameters close to one, and a large batch size. Next, we show that there exists a batch size called the critical batch size minimizing the stochastic first-order oracle (SFO) complexity, which is the stochastic gradient computation cost, and that SFO complexity increases once the batch size exceeds the critical batch size. Finally, we provide numerical results that support our theoretical results. That is, the numerical results indicate that Adam using a small constant learning rate, hyperparameters close to one, and the critical batch size minimizing SFO complexity has faster convergence than Momentum and stochastic gradient descent (SGD).
****

Jointist: Simultaneous Improvement of Multi-instrument Transcription and Music Source Separation via Joint Training
Kin Wai Cheuk,Keunwoo Choi,Qiuqiang Kong,Bochen Li,Minz Won,Ju-Chiang Wang,Yun-Ning Hung,Dorien Herremans
In this paper, we introduce Jointist, an instrument-aware multi-instrument frame

work that is capable of transcribing, recognizing, and separating multiple musical instruments from an audio clip.
Jointist consists of an instrument recognition module that conditions the other two modules: a transcription module that outputs instrument-specific piano rolls, and a source separation module that utilizes instrument information and transcription results. The joint training of the transcription and source separation modules serves to improve the performance of both tasks. The instrument module is optional and can be directly controlled by human users. This makes Jointist a flexible user-controllable framework.

Our challenging problem formulation makes the model highly useful in the real world given that modern popular music typically consists of multiple instruments. Its novelty, however, necessitates a new perspective on how to evaluate such a model. In our experiments, we assess the proposed model from various aspects, providing a new evaluation perspective for multi-instrument transcription. Subjective listening test shows that Jointist achieves state-of-the-art performance on popular music, outperforming existing multi-instrument transcription models such as MT3. %We also argue that transcription models can be used as a preprocessing module for other music analysis tasks. We conducted experiments on several downstream tasks, and found that the proposed method improved transcription by more than 1 percentage points (ppt.); source separation by 5 SDR, downbeat detection by 1.8 ppt., chord recognition by 1.4 ppt., and key estimation by 1.4 ppt., when utilizing transcription results obtained from Jointist.
****
Where prior learning can and can't work in unsupervised inverse problems
Benoît Malézieux,Florent Michel,Matthieu Kowalski,Thomas Moreau
Linear inverse problems consist in recovering a signal from its noisy observation in a lower dimensional space. Many popular resolution methods rely on data-driven algorithms that learn a prior from pairs of signals and observations to overcome the loss of information. However, these approaches are difficult, if not impossible, to adapt to unsupervised contexts -- where no ground truth data are available -- due to the need for learning from clean signals. This paper studies situations that do or do not allow learning a prior in unsupervised inverse problems. First, we focus on dictionary learning and point out that recovering the dictionary is unfeasible without constraints when the signal is observed through only one measurement operator. It can, however, be learned with multiple operators, given that they are diverse enough to span the whole signal space. Then, we study methods where weak priors are made available either through optimization constraints or deep learning architectures. We empirically emphasize that they perform better than hand-crafted priors only if they are adapted to the inverse problem.
****
When are smooth-ReLUs ReLU-like?
Ermal Rrapaj,Luca Celotti,Qiyao Wei,Martin Magill
ReLU is one of the most popular activations in deep learning, especially thanks to its stabilizing effect on training. However, because it is non-differentiable at the origin, it complicates the use of analysis methods that examine derivatives, such as the Neural Tangent Kernel (NTK). Many smooth relaxations try to retain the practical benefits of ReLU while increasing network regularity. Although their success has ranged widely, some notable architectures (e.g., the BERT family) do utilize them. We present a theoretical characterization of smooth-ReLUs within fully-connected feed-forward neural networks. In addition to the well-known SWISH and GeLU, we introduce GumbelLU, AlgebraicLU, and GudermanLU, as new relaxations. All these activations can be characterized by a positive temperature parameter which we can lower to continuously improve the approximation. By studying the interplay of initialization schemes with temperature, we confirm that when these relaxations converge uniformly to ReLU, the statistical properties of the corresponding neural networks at initialization also converge to those of ReLU networks. Moreover, we derive temperature-dependent critical initialization schemes with which networks based on these activations exhibit stable ReLU-like

behavior at any temperature. Finally, we empirically study both classes of networks on MNIST and CIFAR-10 in the full-batch training regime. We show that, while all networks exhibit very similar train loss trajectories at criticality, smooth-ReLU networks feature differentiable NTKs throughout training, whereas ReLU networks exhibit stochastic NTK fluctuations. Our results clarify how smooth-ReLU relaxations reproduce the practical benefits of ReLU in everywhere-smooth neural networks.

****

SpectraNet: multivariate forecasting and imputation under distribution shifts and missing data

Cristian Ignacio Challu,Peihong Jiang,Ying Nian Wu,Laurent Callot

In this work, we tackle two widespread challenges in real applications for time-series forecasting that have been largely understudied: distribution shifts and missing data. We propose SpectraNet, a novel multivariate time-series forecasting model that dynamically infers a latent space spectral decomposition to capture current temporal dynamics and correlations on the recent observed history. A Convolution Neural Network maps the learned representation by sequentially mixing its components and refining the output. Our proposed approach can simultaneously produce forecasts and interpolate past observations and can, therefore, greatly simplify production systems by unifying imputation and forecasting tasks into a single model. SpectraNet achieves SoTA performance simultaneously on both tasks on five benchmark datasets, compared to forecasting and imputation models, with up to 92% fewer parameters and comparable training times. On settings with up to 80% missing data, SpectraNet has average performance improvements of almost 50% over the second-best alternative.

****

An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems

Andrea Gesmundo,Jeff Dean

Multitask learning assumes that models capable of learning from multiple tasks can achieve better quality and efficiency via knowledge transfer, a key feature of human learning. Though, state of the art ML models rely on high customization for each task and leverage size and data scale rather than scaling the number of tasks. Also, continual learning, that adds the temporal aspect to multitask, is often focused to the study of common pitfalls such as catastrophic forgetting instead of being studied at a large scale as a critical component to build the next generation artificial intelligence.

We propose an evolutionary method capable of generating large scale multitask models that support the dynamic addition of new tasks. The generated multitask models are sparsely activated and integrates a task-based routing that guarantees bounded compute cost and fewer added parameters per task as the model expands. The proposed method relies on a knowledge compartmentalization technique to achieve immunity against catastrophic forgetting and other common pitfalls such as gradient interference and negative transfer. We demonstrate empirically that the proposed method can jointly solve and achieve competitive results on 69 public image classification tasks, for example improving the state of the art on a competitive benchmark such as cifar10 by achieving a 15% relative error reduction compared to the best model trained on public data.

****

Uncertainty and Traffic Light Aware Pedestrian Crossing Intention Prediction

Minali Upreti,Jayanth Ramesh,Chandan R Kumar,Bodhisattwa Chakraborty,VIKRAM BALISAVIRA,Phillip Czech,Vitali Kaiser,Markus Roth

Predicting Vulnerable Road User (VRU) crossing intention is one of the major challenges in automated driving. Crossing intention prediction systems trained only on pedestrian features underperform in situations that are most obvious to humans, as the latter take additional context features into consideration. Moreover, such systems tend to be over-confident for out-of-distribution samples, therefore making them less reliable to be used by downstream tasks like sensor fusion and trajectory planning for automated vehicles. In this work, we demonstrate that the results of crossing intention prediction systems can be improved by incorpo

rating traffic light status as an additional input. Further, we make the model robust and interpretable by estimating uncertainty. Experiments on the PIE dataset show that the F1-score improved from 0.77 to 0.82 and above for three different baseline systems when considering traffic-light context. By adding uncertainty, we show increased uncertainty values for out-of-distribution samples, therefore leading to interpretable and reliable predictions of crossing intention.
****

Benchmarking Constraint Inference in Inverse Reinforcement Learning
Guiliang Liu,Yudong Luo,Ashish Gaurav,Kasra Rezaee,Pascal Poupart

When deploying Reinforcement Learning (RL) agents into a physical system, we must ensure that these agents are well aware of the underlying constraints. In many real-world problems, however, the constraints are often hard to specify mathematically and unknown to the RL agents. To tackle these issues, Inverse Constrained Reinforcement Learning (ICRL) empirically estimates constraints from expert demonstrations. As an emerging research topic, ICRL does not have common benchmarks, and previous works tested algorithms under hand-crafted environments with manually-generated expert demonstrations. In this paper, we construct an ICRL benchmark in the context of RL application domains, including robot control, and autonomous driving. For each environment, we design relevant constraints and train expert agents to generate demonstration data. Besides, unlike existing baselines that learn a deterministic constraint, we propose a variational ICRL method to model a posterior distribution of candidate constraints. We conduct extensive experiments on these algorithms under our benchmark and show how they can facilitate studying important research challenges for ICRL. The benchmark, including the instructions for reproducing ICRL algorithms, is available at https://github.com/Guiliang/ICRL-benchmarks-public.
****

Forward and Backward Lifelong Learning with Time-dependent Tasks
Veronica Alvarez,Santiago Mazuelas,Jose A. Lozano

For a sequence of classification tasks that arrive over time, lifelong learning methods can boost the effective sample size of each task by leveraging information from preceding and succeeding tasks (forward and backward learning). However, backward learning is often prone to a so-called catastrophic forgetting in which a task’s performance gets worse while trying to repeatedly incorporate information from succeeding tasks. In addition, current lifelong learning techniques are designed for i.i.d. tasks and cannot capture the usual higher similarities between consecutive tasks. This paper presents lifelong learning methods based on minimax risk classifiers (LMRCs) that effectively exploit forward and backward learning and account for time-dependent tasks. In addition, we analytically characterize the increase in effective sample size provided by forward and backward learning in terms of the tasks’ expected quadratic change. The experimental evaluation shows that LMRCs can result in a significant performance improvement, especially for reduced sample sizes.
****

Memory Gym: Partially Observable Challenges to Memory-Based Agents
Marco Pleines,Matthias Pallasch,Frank Zimmer,Mike Preuss

Memory Gym is a novel benchmark for challenging Deep Reinforcement Learning agents to memorize events across long sequences, be robust to noise, and generalize. It consists of the partially observable 2D and discrete control environments Mortar Mayhem, Mystery Path, and Searing Spotlights. These environments are believed to be unsolvable by memory-less agents because they feature strong dependencies on memory and frequent agent-memory interactions. Empirical results based on Proximal Policy Optimization (PPO) and Gated Recurrent Unit (GRU) underline the strong memory dependency of the contributed environments. The hardness of these environments can be smoothly scaled, while different levels of difficulty (some of them unsolved yet) emerge for Mortar Mayhem and Mystery Path. Surprisingly, Searing Spotlights poses a tremendous challenge to GRU-PPO, which remains an open puzzle. Even though the
randomly moving spotlights reveal parts of the environment’s ground truth, environmental ablations hint that these pose a severe perturbation to agents that lev

erage recurrent model architectures as their memory.
Source Code: https://github.com/MarcoMeter/drl-memory-gym/

****

Worst-case Few-shot Evaluation: Are Neural Networks Robust Few-shot Learners?

Yudong Wang,Ma Chang,Qingxiu Dong,Lingpeng Kong,Zhifang Sui,Jingjing Xu

Neural networks have achieved remarkable performance on various few-shot tasks. However, recent studies reveal that existing few-shot models often exploit the spurious correlations between training and test sets, achieving a high performance that is hard to generalize. Motivated that a robust few-shot learner should accurately classify data given any valid training set, we consider a worst-case few-shot evaluation that computes worst-case generalization errors by constructing a challenging few-shot set. Specifically, we search for the label-balanced subset of a full-size training set that results in the largest expected risks. Since the search space is enormous, we propose an efficient method NMMD-attack to optimize the target by maximizing NMMD distance (maximum mean discrepancy based on neural tangent kernel). Experiments show that NMMD-attack can successfully attack various architectures. The large gap between average performance and worst-case performance shows that neural networks still suffer from poor robustness. We appeal to more worst-case benchmarks for better robust few-shot evaluation.

****

Discovering Policies with DOMiNO: Diversity Optimization Maintaining Near Optimality

Tom Zahavy,Yannick Schroecker,Feryal Behbahani,Kate Baumli,Sebastian Flennerhag,Shaobo Hou,Satinder Singh

In this work we propose a Reinforcement Learning (RL) agent that can discover complex behaviours in a rich environment with a simple reward function. We define diversity in terms of state-action occupancy measures, since policies with different occupancy measures visit different states on average. More importantly, defining diversity in this way allows us to derive an intrinsic reward function for maximizing the diversity directly. Our agent, DOMiNO, stands for Diversity Optimization Maintaining Near Optimally. It is based on maximizing a reward function with two components: the extrinsic reward and the diversity intrinsic reward, which are combined with Lagrange multipliers to balance the quality-diversity trade-off. Any RL algorithm can be used to maximize this reward and no other changes are needed. We demonstrate that given a simple reward functions in various control domains, like height (stand) and forward velocity (walk), DOMiNO discovers diverse and meaningful behaviours. We also perform extensive analysis of our approach, compare it with other multi-objective baselines, demonstrate that we can control both the quality and the diversity of the set via interpretable hyperparameters, and show that the set is robust to perturbations of the environment.

****

SpeedyZero: Mastering Atari with Limited Data and Time

Yixuan Mei,Jiaxuan Gao,Weirui Ye,Shaohuai Liu,Yang Gao,Yi Wu

Many recent breakthroughs of deep reinforcement learning (RL) are mainly built upon large-scale distributed training of model-free methods using millions to billions of samples. On the other hand, state-of-the-art model-based RL methods can achieve human-level sample efficiency but often take a much longer over all training time than model-free methods. However, high sample efficiency and fast training time are both important to many real-world applications. We develop SpeedyZero, a distributed RL system built upon a state-of-the-art model-based RL method, EfficientZero, with a dedicated system design for fast distributed computation. We also develop two novel algorithmic techniques, Priority Refresh and Clipped LARS, to stabilize training with massive parallelization and large batch size. SpeedyZero maintains on-par sample efficiency compared with EfficientZero while achieving a 14.5X speedup in wall-clock time, leading to human-level performances on the Atari benchmark within 35 minutes using only 300k samples. In addition, we also present an in-depth analysis on the fundamental challenges in further scaling our system to bring insights to the community.

****

HT-Net: Hierarchical Transformer based Operator Learning Model for Multiscale P

DEs
Xinliang Liu,Bo Xu,Lei Zhang
Complex nonlinear interplays of multiple scales give rise to many interesting physical phenomena and pose major difficulties for the computer simulation of multiscale PDE models in areas such as reservoir simulation, high frequency scattering and turbulence modeling. In this paper, we introduce a hierarchical transformer (HT-Net) scheme to efficiently learn the solution operator for multiscale PDEs. We construct a hierarchical architecture with scale adaptive interaction range, such that the features can be computed in a nested manner and with a controllable linear cost. Self-attentions over a hierarchy of levels can be used to encode and decode the multiscale solution space over all scale ranges. In addition, we adopt an empirical H^1 loss function to counteract the spectral bias of the neural network approximation for multiscale functions. In the numerical experiments, we demonstrate the superior performance of the HT-Net scheme compared with state-of-the-art (SOTA) methods for representative multiscale problems.

****
Multi-Agent Multi-Game Entity Transformer
Rundong Wang,Weixuan Wang,Xianhan Zeng,Liang Wang,Zhenjie Lian,Yiming Gao,Feiyu Liu,Siqin Li,Xianliang Wang,QIANG FU,Yang Wei,Lanxiao Huang,Longtao Zheng,Zinovi Rabinovich,Bo An
Building large-scale generalist pre-trained models for many tasks is becoming an emerging and potential direction in reinforcement learning (RL). Research such as Gato and Multi-Game Decision Transformer have displayed outstanding performance and generalization capabilities on many games and domains. However, there exists a research blank about developing highly capable and generalist models in multi-agent RL (MARL), which can substantially accelerate progress towards general AI. To fill this gap, we propose Multi-Agent multi-Game ENtity TrAnsformer (MAGENTA) from the entity perspective as an orthogonal research to previous time-sequential modeling. Specifically, to deal with different state/observation spaces in different games, we analogize games as languages, thus training different "tokenizers" for various games. The feature inputs are split according to different entities and tokenized in the same continuous space. Then, two types of transformer-based model are proposed as permutation-invariant architectures to deal with various numbers of entities and capture the attention over different entities. MAGENTA is trained on Honor of Kings, Starcraft II micromanagement, and Neural MMO with a single set of transformer weights. Extensive experiments show that MAGENTA can play games across various categories with arbitrary numbers of agents and increase the efficiency of fine-tuning in new games and scenarios by 50\%-100\%. See our project page at \url{https://sites.google.com/view/rl-magenta}.
****
On the Convergence of Gradient Flow on Multi-layer Linear Models
Hancheng Min,Rene Vidal,Enrique Mallada
In this paper, we analyze the convergence of gradient flow on a multi-layer linear model with a loss function of the form $f(W_1W_2\cdots W_L)$. We show that when f satisfies the gradient dominance property, proper weight initialization leads to exponential convergence of the gradient flow to a global minimum of the loss. Moreover, the convergence rate depends on two trajectory-specific quantities that are controlled by the weight initialization: the \emph{imbalance matrices}, which measure the difference between the weights of adjacent layers, and the least singular value of the \emph{weight product} $W=W_1W_2\cdots W_L$. Our analysis provides improved rate bounds for several multi-layer network models studied in the literature, leading to novel characterizations of the effect of weight imbalance on the rate of convergence. Our results apply to most regression losses and extend to classification ones.
****
Neural Architecture Design and Robustness: A Dataset
Steffen Jung,Jovita Lukasik,Margret Keuper
Deep learning models have proven to be successful in a wide range of machine learning tasks. Yet, they are often highly sensitive to perturbations on the input

data which can lead to incorrect decisions with high confidence, hampering their deployment for practical use-cases. Thus, finding architectures that are (more) robust against perturbations has received much attention in recent years. Just like the search for well-performing architectures in terms of clean accuracy, this is usually involves a tedious trial-and-error process with one additional challenge: the evaluation of a network's robustness is significantly more expensive than its evaluation for clean accuracy. Thus, the aim of this paper is to facilitate better streamlined research on architectural design choices with respect to their impact on robustness as well as, for example, the evaluation of surrogate measures for robustness. We therefore borrow one of the most commonly considered search spaces for neural architecture search for image classification, NAS-Bench-201, which contains a manageable size of 6466 non-isomorphic network designs. We evaluate all these networks on a range of common adversarial attacks and corruption types and introduce a database on neural architecture design and robustness evaluations. We further present three exemplary use cases of this dataset, in which we (i) benchmark robustness measurements based on Jacobian and Hessian matrices for their robustness predictability, (ii) perform neural architecture search on robust accuracies, and (iii) provide an initial analysis of how architectural design choices affect robustness. We find that carefully crafting the topology of a network can have substantial impact on its robustness, where networks with the same parameter count range in mean adversarial robust accuracy from 20%-41%. Code and data is available at http://robustness.vision/.
****

Does Deep Learning Learn to Abstract? A Systematic Probing Framework

Shengnan An,Zeqi Lin,Bei Chen,Qiang Fu,Nanning Zheng,Jian-Guang Lou

Abstraction is a desirable capability for deep learning models, which means to induce abstract concepts from concrete instances and flexibly apply them beyond the learning context. At the same time, there is a lack of clear understanding about both the presence and further characteristics of this capability in deep learning models. In this paper, we introduce a systematic probing framework to explore the abstraction capability of deep learning models from a transferability perspective. A set of controlled experiments are conducted based on this framework, providing strong evidence that two probed pre-trained language models (PLMs), T5 and GPT2, have the abstraction capability. We also conduct in-depth analysis, thus shedding further light: (1) the whole training phase exhibits a "memorize-then-abstract" two-stage process; (2) the learned abstract concepts are gathered in a few middle-layer attention heads, rather than being evenly distributed throughout the model; (3) the probed abstraction capabilities exhibit robustness against concept mutations, and are more robust to low-level/source-side mutations than high-level/target-side ones; (4) generic pre-training is critical to the emergence of abstraction capability, and PLMs exhibit better abstraction with larger model sizes and data scales.
****

Learning to mine approximate network motifs

Carlos Oliver,Dexiong Chen,Vincent Mallet,Pericles Philippopoulos,Karsten Borgwardt

Frequent and structurally related subgraphs, also known as network motifs, are valuable features of many datasets. However, strong combinatorial bottlenecks have made it difficult to extract motifs and use them in learning tasks without strong constraints on the motif properties. In this work we propose a representation learning method based on learnable graph coarsening, MotiFiesta which is the first to be able to extract large and approximate motifs in a fully differentiable manner. We build benchmark datasets and evaluation metrics which test the ability our proposed and future models to capture different aspects of motif discovery where ground truth motifs are not known. Finally, explore the notion of exploiting learned motifs as an inductive bias in real-world datasets by showing competitive performance on motif-based featuresets with established real-world benchmark datasets against concurrent architectures.
****

Automatic Clipping: Differentially Private Deep Learning Made Easier and Stronge r

Zhiqi Bu,Yu-Xiang Wang,Sheng Zha,George Karypis

Per-example gradient clipping is a key algorithmic step that enables practical differential private (DP) training for deep learning models. The choice of clipping threshold R, however, is shown to be vital for achieving high accuracy under DP. We propose an easy-to-use replacement, called automatic clipping, that eliminates the need to tune R for any DP optimizers, including DP-SGD, DP-Adam, DP-LAMB and many others.

The automatic variants are as private and computationally efficient as existing DP optimizers, but require no DP-specific hyperparameters and thus make DP training as amenable as the standard non-private training. We give a rigorous convergence analysis of automatic DP-SGD in the non-convex setting, which shows that it {can enjoy an asymptotic convergence rate that matches the standard SGD, under a symmetric gradient noise assumption of the per-sample gradients.} We also demonstrate on various language and vision tasks that automatic clipping outperforms or matches the state-of-the-art, and can be easily employed with minimal changes to existing codebases.

****

Trust Your ∇: Gradient-based Intervention Targeting for Causal Discovery

Mateusz Olko,Micha■ Zaj■c,Aleksandra Nowak,Nino Scherrer,Yashas Annadani,Stefan Bauer,■ukasz Kuci■ski,Piotr Mi■o■

Inferring causal structure from data is a challenging task of fundamental importance in science. Observational data are often insufficient to identify a system’s causal structure uniquely. While conducting interventions (i.e., experiments) can improve the identifiability, such samples are usually challenging and expensive to obtain. Hence, experimental design approaches for causal discovery aim to minimize the number of interventions by estimating the most informative intervention target. In this work, we propose a novel gradient-based intervention targeting method, abbreviated GIT, that 'trusts' the gradient estimator of a gradient-based causal discovery framework to provide signals for intervention acquisition function. We provide extensive experiments in simulated and real-world datasets and demonstrate that GIT performs on par with competitive baselines, surpassing them in the low-data regime.

****

Improving Out-of-distribution Generalization with Indirection Representations

Kha Pham,Hung Le,Man Ngo,Truyen Tran

We propose a generic module named Indirection Layer (InLay), which leverages indirection and data internal relationships to effectively construct symbolic indirect representations to improve out-of-distribution generalization capabilities of various neural architectures. InLay receives data input in the form of a sequence of objects, treats it as a complete weighted graph whose vertices are the objects and edge weights are scalars representing relationships between vertices. The input is first mapped via indirection to a symbolic graph with data-independent and trainable vertices. This symbolic graph is then propagated, resulting in new vertex features whose indirection will be used for prediction steps afterward. Theoretically, we show that the distances between indirection representations are bounded by the distances between corresponding graphs, implying that unseen samples with very different surface statistics can still be close in the representation space to the seen samples if they share similar internal relationships. We demonstrate that InLay is consistently effective in improving out-of-distribution generalization throughout a comprehensive suite of experiments, including IQ problems, distorted image classification, and few-shot domain adaptation NLP classification. We also conduct ablation studies to verify different design choices of InLay.

****

Accelerating Guided Diffusion Sampling with Splitting Numerical Methods

Suttisak Wizadwongsa,Supasorn Suwajanakorn

Guided diffusion is a technique for conditioning the output of a diffusion model at sampling time without retraining the network for each specific task. However, one drawback of diffusion models, whether they are guided or unguided, is thei r slow sampling process.

Recent techniques can accelerate unguided sampling by applying high-order numerical methods to the sampling process when viewed as differential equations. On the contrary, we discover that the same techniques do not work for guided sampling, and little has been explored about its acceleration. This paper explores the culprit of this problem and provides a solution based on operator splitting methods, motivated by our key finding that classical high-order numerical methods are unsuitable for the conditional function. Our proposed method can re-utilize the high-order methods for guided sampling and can generate images with the same quality as a 250-step DDIM baseline using 32-58% less sampling time on ImageNet256.

We also demonstrate usage on a wide variety of conditional generation tasks, such as text-to-image generation, colorization, inpainting, and super-resolution.

****

RealSinger: Ultra-Realistic Singing Voice Generation via Stochastic Differential Equations

Shoule Wu,Ziqiang Shi

Synthesizing high-quality singing voice from music score is a challenging problem in music generation and has many practical applications. Samples generated by existing singing voice synthesis (SVS) systems can roughly reflect the lyrics, pitch and duration in a given score, but they fail to contain necessary details. In this paper, based on stochastic differential equations (SDE) we propose RealSinger to generate 22.05kHz ultra-realistic singing voice conditioned on a music score. Our RealSinger learns to find the stochastic process path from a source of white noise to the target singing voice manifold under the conditional music score, allowing to sing the music score while maintaining the local voice details of the target singer. During training, our model learns to accurately predict the direction of movement in the ambient Euclidean space onto the low-dimensional singing voice manifold. RealSinger's framework is very flexible. It can either generate intermediate feature representations of the singing voice, such as mel-spectrogram, or directly generate the final waveform, as in the end-to-end style which rectify defects and accumulation errors introduced by two-stage connected singing synthesis systems. An extensive subjective and objective test on benchmark datasets shows significant gains in perceptual quality using RealSinger. The mean opinion scores (MOS) obtained with RealSinger are closer to those of the human singer's original high-fidelity singing voice than to those obtained with any state-of-the-art method. Audio samples are available at https://realsinger.github.io/.

****

Demystifying Approximate RL with ϵ-greedy Exploration: A Differential Inclusion View

Aditya Gopalan,Gugan Thoppe

Q-learning and SARSA(0) with ϵ-greedy exploration are leading reinforcement learning methods, and their tabular forms converge to the optimal Q-function under reasonable conditions. However, with function approximation, they exhibit unexpected behaviors, such as i.) policy oscillation and chattering, and ii.) convergence to different attractors (possibly even the worst policy) on different runs, ii.) multiple attractors, and iii.) worst policy convergence, apart from the textbook instability. Accordingly, a theory to explain these phenomena has been a long-standing open problem, even for basic linear function approximation (Sutton, 1999). Our work uses differential inclusion theory to provide the first framework for resolving this problem. We further illustrate via numerical examples how this framework helps explain these algorithms' asymptotic behaviors.

****

Batch Multivalid Conformal Prediction

Christopher Jung,Georgy Noarov,Ramya Ramalingam,Aaron Roth

We develop fast distribution-free conformal prediction algorithms for obtaining multivalid coverage on exchangeable data in the batch setting. Multivalid coverage guarantees are stronger than marginal coverage guarantees in two ways: (1) They hold even conditional on group membership---that is, the target coverage lev el $1-\alpha$ holds conditionally on membership in each of an arbitrary (potentially intersecting) group in a finite collection $\mathcal{G}$ of regions in the feature space. (2) They hold even conditional on the value of the threshold used to produce the prediction set on a given example. In fact multivalid coverage guarantees hold even when conditioning on group membership and threshold value simultaneously.

We give two algorithms: both take as input an arbitrary non-conformity score and an arbitrary collection of possibly intersecting groups $\mathcal{G}$, and then can equip arbitrary black-box predictors with prediction sets. Our first algorithm is a direct extension of quantile regression, needs to solve only a single convex minimization problem, and produces an estimator which has group-conditional guarantees for each group in $\mathcal{G}$. Our second algorithm is iterative, and gives the full guarantees of multivalid conformal prediction: prediction sets that are valid conditionally both on group membership and non-conformity threshold. We evaluate the performance of both of our algorithms in an extensive set of experiments.

****

Transferring Pretrained Diffusion Probabilistic Models

Fuming You,Zhou Zhao

Diffusion Probabilistic Models (DPMs) achieve impressive performance in visual generative tasks recently. However, the success of DPMs heavily relies on large amounts of data and optimization steps, which limits the application of DPMs to small datasets and limited computational resources. In this paper, we investigate transfer learning in DPMs to leverage the DPMs pretrained on large-scale datasets for generation with limited data. Firstly, we show that previous methods like training from scratch or determining the transferable parts is not suitable for the DPM due to its U-Net based denoising architecture with the external denoising timestep input. To address it, we present a condition-based tuning approach to take full advantages of existing pretrained models. Concretely, we obtain the semantic embeddings of condition images by the pretrained CLIP model, and then inject these semantic informations to the pretrained DPM via a ''Attention-NonLinear'' (ANL) module. The adaptation to a new task can be achieved by only tuning the ANL module inserted into the pretrained DPM hierarchically. To further enhance the diversity of generated images, we introduce a masked sampling strategy based on the condition mechanism. Extensive experiments validate the effectiveness and efficiency of our proposed tuning approach in generative task transfer and data augmentation for semi-supervised learning.

****

ELBO-ing Stein Mixtures

Ola Rønning,Christophe Ley,Ahmad Salim Al-Sibahi,Thomas Hamelryck

Stein variational gradient descent (SVGD) \citep{DBLP:conf/nips/LiuW16} is a particle-based technique for Bayesian inference. SVGD has recently gained popularity because it combines the ability of variational inference to handle tall data with the modeling power of non-parametric inference. Unfortunately, the number of particles required to represent a model adequately grows exponentially with the dimensionality of the model. Stein mixtures \citep{nalisnick2017variational} alleviate the exponential growth in particles by letting each particle parameterize a distribution. However, the inference algorithm proposed by \cite{nalisnick2017variational} can be numerically unstable. We show that their algorithm corresponds to inference with the R\'enyi α-divergence for $\alpha=0$ and that using other values for α can lead to more stable inference. We empirically study the performance of Stein mixtures inferred with different α values on various real-world problems, demonstrating significantly improved results when using $\alpha=1$, which coincides with using the evidence lower bound (ELBO). We call this instance of our algorithm ELBO-within-Stein. A black-box version of the inference algorithm (for arbitrary $\alpha\in \sR$) is available in the deep probabilistic programming language NumPyro \citep{phan2019}.

****

Schedule-Robust Online Continual Learning

Ruohan Wang,Marco Ciccone,Giulia Luise,Andrew Yapp,massimiliano pontil,Carlo Ciliberto

A continual learning (CL) algorithm learns from a non-stationary data stream. The non-stationarity is modeled by some schedule that determines how data is presented over time. Most current methods make strong assumptions on the schedule and have unpredictable performance when such requirements are not met. A key challenge in CL is thus to design methods robust against arbitrary schedules over the same underlying data, since in real-world scenarios schedules are often unknown and dynamic. In this work, we introduce the notion of schedule-robustness for CL and a novel approach satisfying this desirable property in the challenging online class-incremental setting. We also present a new perspective on CL, as the process of learning a schedule-robust predictor, followed by adapting the predictor using only replay data. Empirically, we demonstrate that our approach outperforms existing methods on CL benchmarks for image classification by a large margin.

****

On the Usefulness of Embeddings, Clusters and Strings for Text Generation Evaluation

Tiago Pimentel,Clara Isabel Meister,Ryan Cotterell

A good automatic evaluation metric for language generation ideally correlates highly with human judgements of text quality. Yet, there is a dearth of such metrics, which inhibits the rapid and efficient progress of language generators. One exception is the recently proposed Mauve. In theory, Mauve measures an information-theoretic divergence between two probability distributions over strings: one representing the language generator under evaluation; the other representing the true natural language distribution. Mauve's authors argue that its success comes from the qualitative properties of their proposed divergence. Yet in practice, as this divergence is uncomputable, Mauve approximates it by measuring the divergence between multinomial distributions over clusters instead, where cluster assignments are attained by grouping strings based on a pretrained language model's embeddings. As we show, however, this is not a tight approximation---in either theory or practice. This begs the question: why does Mauve work so well? In this work, we show that \mauve was right for the wrong reasons, and that its newly proposed divergence is not necessary for its high performance. In fact, classical divergences paired with its proposed cluster-based approximation may actually serve as better evaluation metrics. We finish the paper with a probing analysis; this analysis leads us to conclude that---by encoding syntactic- and coherence-level features of text, while ignoring surface-level features---such cluster-based approximations to string distributions may simply be better for evaluating state-of-the-art language generators.

****

Attention Enables Zero Approximation Error

Zhiying Fang,Yidong Ouyang,Ding-Xuan Zhou,Guang Cheng

Attention-based architectures become the core backbone of many state-of-the-art models for various tasks, including language translation and image classification. However, theoretical properties of attention-based models are seldom considered. In this work, we show that with suitable adaptations, the single-head self-attention transformer with a fixed number of transformer encoder blocks and free parameters is able to generate any desired polynomial of the input with no error. The number of transformer encoder blocks is the same as the degree of the target polynomial. Even more exciting, we find that these transformer encoder blocks in this model do not need to be trained. As a direct consequence, we show that the single-head self-attention transformer with increasing numbers of free parameters is universal. Also, we show that our proposed model can avoid the classical trade-off between approximation error and sample error in the mean squared error analysis for the regression task if the target function is a polynomial. We conduct various experiments and ablation studies to verify our theoretical results.

****

Revisiting Activation Function Design for Improving Adversarial Robustness at Sc

ale
Cihang Xie,Mingxing Tan,Boqing Gong,Alan Yuille,Quoc V Le

Modern ConvNets typically use ReLU activation function. Recently smooth activation functions have been used to improve their accuracy. Here we study the role of smooth activation function from the perspective of adversarial robustness. We find that ReLU activation function significantly weakens adversarial training due to its non-smooth nature. Replacing ReLU with its smooth alternatives allows adversarial training to find harder adversarial training examples and to compute better gradient updates for network optimization.

We focus our study on the large-scale ImageNet dataset. On ResNet-50, switching from ReLU to the smooth activation function SILU improves adversarial robustness from 33.0% to 42.3%, while also improving accuracy by 0.9% on ImageNet. Smooth activation functions also scale well with larger networks: it helps EfficientNet-L1 to achieve 82.2% accuracy and 58.6% robustness, largely outperforming the previous state-of-the-art defense by 9.5% for accuracy and 11.6% for robustness. Models are available at https://rb.gy/qt8jya.

****

Contrastive Hierarchical Clustering
Micha■ Znale■niak,Przemys■aw Rola,Patryk Kaszuba,Jacek Tabor,Marek ■mieja

Deep clustering has been dominated by flat clustering models, which split a dataset into a predefined number of groups. Although recent methods achieve extremely high similarity with the ground truth on popular benchmarks, the information contained in the flat partition is limited. In this paper, we introduce CoHiClust, a Contrastive Hierarchical Clustering model based on deep neural networks, which can be applied to large-scale image data. By employing a self-supervised learning approach, CoHiClust distills the base network into a binary tree without access to any labeled data. The hierarchical clustering structure can be used to analyze the relationship between clusters as well as to measure the similarity between data points. Experiments performed on typical image benchmarks demonstrate that CoHiClust generates a reasonable structure of clusters, which is consistent with our intuition and image semantics. Moreover, by applying the proposed pruning strategy, we can restrict the hierarchy to the requested number of clusters (leaf nodes) and obtain the clustering accuracy comparable to the state-of-the-art flat clustering baselines.

****

Accurate Bayesian Meta-Learning by Accurate Task Posterior Inference
Michael Volpp,Philipp Dahlinger,Philipp Becker,Christian Daniel,Gerhard Neumann

Bayesian meta-learning (BML) enables fitting expressive generative models to small datasets by incorporating inductive priors learned from a set of related tasks. The Neural Process (NP) is a prominent deep neural network-based BML architecture, which has shown remarkable results in recent years. In its standard formulation, the NP encodes epistemic uncertainty in an amortized, factorized, Gaussian variational (VI) approximation to the BML task posterior (TP), using reparametrized gradients. Prior work studies a range of architectural modifications to boost performance, such as attentive computation paths or improved context aggregation schemes, while the influence of the VI scheme remains under-explored. We aim to bridge this gap by introducing GMM-NP, a novel BML model, which builds on recent work that enables highly accurate, full-covariance Gaussian mixture (GMM) TP approximations by combining VI with natural gradients and trust regions. We show that GMM-NP yields tighter evidence lower bounds, which increases the efficiency of marginal likelihood optimization, leading to improved epistemic uncertainty estimation and accuracy. GMM-NP does not require complex architectural modifications, resulting in a powerful, yet conceptually simple BML model, which outperforms the state of the art on a range of challenging experiments, highlighting its applicability to settings where data is scarce.

****

Learning to Decompose Visual Features with Latent Textual Prompts
Feng Wang,Manling Li,Xudong Lin,Hairong Lv,Alex Schwing,Heng Ji

Recent advances in pre-training vision-language models like CLIP have shown great potential in learning transferable visual representations. Nonetheless, for downstream inference, CLIP-like models suffer from either 1) degraded accuracy and robustness in the case of inaccurate text descriptions during retrieval-based inference (the challenge for zero-shot protocol); or 2) breaking the well-established vision-language alignment (the challenge for linear probing). To address them, we propose Decomposed Feature Prompting (DeFo). DeFo leverages a flexible number of learnable embeddings as textual input while maintaining the vision-language dual-model architecture, which enables the model to learn decomposed visual features with the help of feature-level textual prompts. We further use an additional linear layer to perform classification, allowing a scalable size of language inputs. Our empirical study shows DeFo's significance in improving the vision-language models. For example, DeFo obtains 73.2% test accuracy on ImageNet with a ResNet-50 backbone without tuning any pretrained weights of both the vision and language encoder, outperforming zero-shot CLIP by a large margin of 15.0%, and outperforming state-of-the-art vision-language prompt tuning method by 7.6%.

****

Skill Machines: Temporal Logic Composition in Reinforcement Learning

Geraud Nangue Tasse,Devon Jarvis,Steven James,Benjamin Rosman

A major challenge in reinforcement learning is specifying tasks in a manner that is both interpretable and verifiable. One common approach is to specify tasks through reward machines---finite state machines that encode the task to be solved. We introduce skill machines, a representation that can be learned directly from these reward machines that encode the solution to such tasks. We propose a framework where an agent first learns a set of base skills in a reward-free setting, and then combines these skills with the learned skill machine to produce composite behaviours specified by any regular language, such as linear temporal logics. This provides the agent with the ability to map from complex logical task specifications to near-optimal behaviours zero-shot. We demonstrate our approach in both a tabular and high-dimensional video game environment, where an agent is faced with several of these complex, long-horizon tasks. Our results indicate that the agent is capable of satisfying extremely complex task specifications, producing near optimal performance with no further learning. Finally, we demonstrate that the performance of skill machines can be improved with regular off-policy reinforcement learning algorithms when optimal behaviours are desired.

****

Surrogate Gradient Design for LIF networks

Luca Celotti,Jean Rouat

Spiking Neuromorphic Computing uses binary activity to improve Artificial Intelligence energy efficiency. However, the non-smoothness of binary ac■tivity requires approximate gradients, known as Surrogate Gradients (SG), to close the performance gap with Deep Learning. Several SG have been proposed in the literature, but it remains unclear how to determine the best SG for a given task and network. Good performance can be achieved with most SG shapes, after an extensive search of hyper-parameters that can be costly. Thus, we aim at experimentally and theoretically define the best SG across different stress tests, to reduce future need of grid search. Here we first show that the derivative of the fast sigmoid outperforms other SG across tasks and networks, for a wide range of learning rates. Secondly, we focus on the Leaky Integrate and Fire (LIF) spiking neural model, and we show that a SG with low dampening, high sharpness, and low tail fat■ness, systematically leads to higher accuracy. Thirdly, we observe that the Orthogonal initialization leads the LIF to higher accuracy with most SG. Fourthly, we note that high initial firing rates, combined with a sparsity encouraging loss term, can lead to better generalization, depending on the SG shape. Finally, we provide a theoretical solution, inspired by Glorot and He initializations, to reduce the need of extensive grid-search, to find an SG and initialization that experimentally result in improved accuracy.

****

Context-enriched molecule representations improve few-shot drug discovery

Johannes Schimunek,Philipp Seidl,Lukas Friedrich,Daniel Kuhn,Friedrich Rippmann, Sepp Hochreiter,Günter Klambauer

A central task in computational drug discovery is to construct models from known active molecules to find further promising molecules for subsequent screening. However, typically only very few active molecules are known. Therefore, few-shot learning methods have the potential to improve the effectiveness of this critical phase of the drug discovery process. We introduce a new method for few-shot drug discovery. Its main idea is to enrich a molecule representation by knowledge about known context or reference molecules. Our novel concept for molecule representation enrichment is to associate molecules from both the support set and the query set with a large set of reference (context) molecules through a modern Hopfield network. Intuitively, this enrichment step is analogous to a human expert who would associate a given molecule with familiar molecules whose properties are known. The enrichment step reinforces and amplifies the covariance structure of the data, while simultaneously removing spurious correlations arising from the decoration of molecules. Our approach is compared with other few-shot methods for drug discovery on the FS-Mol benchmark dataset. On FS-Mol, our approach outperforms all compared methods and therefore sets a new state-of-the art for few-shot learning in drug discovery. An ablation study shows that the enrichment step of our method is the key to improve the predictive quality. In a domain shift experiment, we further demonstrate the robustness of our method.

****

The Multiple Subnetwork Hypothesis: Enabling Multidomain Learning by Isolating Task-Specific Subnetworks in Feedforward Neural Networks

Jacob William Renn,Ian Sotnek,Benjamin Harvey,Brian Caffo

Neural networks have seen an explosion of usage and research in the past decade, particularly within the domains of computer vision and natural language processing. However, only recently have advancements in neural networks yielded performance improvements beyond narrow applications and translated to expanded multitask models capable of generalizing across multiple data types and modalities. Simultaneously, it has been shown that neural networks are overparameterized to a high degree, and pruning techniques have proved capable of significantly reducing the number of active weights within the network while largely preserving performance. In this work, we identify a methodology and network representational structure which allows a pruned network to employ previously unused weights to learn subsequent tasks. We employ these methodologies on well-known benchmarking datasets for testing purposes and show that networks trained using our approaches are able to learn multiple tasks, which may be related or unrelated, in parallel or in sequence without sacrificing performance on any task or exhibiting catastrophic forgetting.

****

Delving into the Openness of CLIP

Shuhuai Ren,Lei Li,Xuancheng Ren,Guangxiang Zhao,Xu Sun

Contrastive Language-Image Pre-training (CLIP) has demonstrated great potential in realizing open-vocabulary visual recognition in a matching style, due to its holistic use of natural language supervision that covers unconstrained real-world visual concepts. However, it is, in turn, also difficult to evaluate and analyze the openness of CLIP-like models, since they are in theory open to any vocabulary but the actual accuracy varies. To address the insufficiency of conventional studies on openness, we resort to an incremental perspective and define the extensibility, which essentially approximates the model's ability to deal with new visual concepts, by evaluating openness through vocabulary expansions. Our evaluation based on extensibility shows that CLIP-like models are hardly truly open and their performances degrade as the vocabulary expands to different degrees. Further analysis reveals that the over-estimation of openness is not because CLIP-like models fail to capture the general similarity of image and text features of novel visual concepts, but because of the confusion among competing text features, that is, they are not stable with respect to the vocabulary. In light of this, we propose to improve the openness of CLIP in feature space by enforcing the distinguishability of text features. Our method retrieves relevant texts from the pre-training corpus to enhance prompts for inference, which boosts the extens ibility and stability of CLIP even without fine-tuning.

****

Test-Time Adaptation via Self-Training with Nearest Neighbor Information

Minguk Jang,Sae-Young Chung,Hye Won Chung

Test-time adaptation (TTA) aims to adapt a trained classifier using online unlabeled test data only, without any information related to the training procedure. Most existing TTA methods adapt the trained classifier using the classifier's prediction on the test data as pseudo-label.

However, under test-time domain shift, accuracy of the pseudo labels cannot be guaranteed, and thus the TTA methods often encounter performance degradation at the adapted classifier. To overcome this limitation, we propose a novel test-time adaptation method, called Test-time Adaptation via Self-Training with nearest neighbor information (TAST), which is composed of the following procedures: (1) adds trainable adaptation modules on top of the trained feature extractor; (2) newly defines a pseudo-label distribution for the test data by using the nearest neighbor information; (3) trains these modules only a few times during test time to match the nearest neighbor-based pseudo label distribution and a prototype-based class distribution for the test data; and (4) predicts the label of test data using the average predicted class distribution from these modules. The pseudo-label generation is based on the basic intuition that a test data and its nearest neighbor in the embedding space are likely to share the same label under the domain shift. By utilizing multiple randomly initialized adaptation modules, TAST extracts useful information for the classification of the test data under the domain shift, using the nearest neighbor information. TAST showed better performance than the state-of-the-art TTA methods on two standard benchmark tasks, domain generalization, namely VLCS, PACS, OfficeHome, and TerraIncognita, and image corruption, particularly CIFAR-10/100C.

****

Learning to Counter: Stochastic Feature-based Learning for Diverse Counterfactual Explanations

Vy Vo,Trung Le,Van Nguyen,He Zhao,Edwin V. Bonilla,Reza Haf,Dinh Phung

Interpretable machine learning seeks to understand the reasoning process of complex black-box systems that are long notorious for lack of explainability. One growing interpreting approach is through counterfactual explanations, which go beyond why a system arrives at a certain decision to further provide suggestions on what a user can do to alter the outcome. A counterfactual example must be able to counter the original prediction from the black-box classifier, while also satisfying various constraints for practical applications. These constraints exist at trade-offs between one and another presenting radical challenges to existing works. To this end, we propose a stochastic learning-based framework that effectively balances the counterfactual trade-offs. The framework consists of a generation and a feature selection module with complementary roles: the former aims to model the distribution of valid counterfactuals whereas the latter serves to enforce additional constraints in a way that allows for differentiable training and amortized optimization. We demonstrate the effectiveness of our method in generating actionable and plausible counterfactuals that are more diverse than the existing methods and particularly in a more efficient manner than counterparts of the same capacity.

****

Accurate Neural Training with 4-bit Matrix Multiplications at Standard Formats

Brian Chmiel,Ron Banner,Elad Hoffer,Hilla Ben-Yaacov,Daniel Soudry

Quantization of the weights and activations is one of the main methods to reduce the computational footprint of Deep Neural Networks (DNNs) training. Current methods enable 4-bit quantization of the forward phase. However, this constitutes only a third of the training process. Reducing the computational footprint of the entire training process requires the quantization of the neural gradients, i.e., the loss gradients with respect to the outputs of intermediate neural layers.

Previous works separately showed that accurate 4-bit quantization of the neural gradients needs to (1) be unbiased and (2) have a log scale. However, no previous work aimed to combine both ideas, as we do in this work. Specifically, we examine the importance of having unbiased quantization in quantized neural network training, where to maintain it, and how to combine it with logarithmic quantization. Based on this, we suggest a $\textit{logarithmic unbiased quantization}$ (LUQ) method to quantize all both the forward and backward phase to 4-bit, achieving state-of-the-art results in 4-bit training without overhead. For example, in ResNet50 on ImageNet, we achieved a degradation of 1.1 %. We further improve this to degradation of only 0.32 % after three epochs of high precision fine-tunining, combined with a variance reduction method---where both these methods add overhead comparable to previously suggested methods.
A reference implementation is supplied in the supplementary material.

****

SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient
Max Ryabinin,Tim Dettmers,Michael Diskin,Alexander Borzunov
Many deep learning applications benefit from using large models with billions of parameters. Training these models is notoriously expensive due to the need for specialized HPC clusters. In this work, we consider alternative setups for training large models: using cheap ``preemptible'' instances or pooling existing resources from multiple regions. We analyze the performance of existing model-parallel algorithms in these conditions and find configurations where training larger models becomes less communication-intensive. Based on these findings, we propose SWARM Parallelism (Stochastically Wired Adaptively Rebalanced Model Parallelism), a model-parallel training algorithm designed for poorly connected, heterogeneous and unreliable devices. SWARM creates temporary randomized pipelines between nodes that are rebalanced in case of failure. We empirically validate our findings and compare SWARM Parallelism with existing large-scale training approaches. Finally, we combine our insights with compression strategies to train a large Transformer language model with 1B shared parameters ($\approx$13B before sharing) on preemptible T4 GPUs with less than 200 Mb/s network.

****

Relative representations enable zero-shot latent space communication
Luca Moschella,Valentino Maiorca,Marco Fumero,Antonio Norelli,Francesco Locatello,Emanuele Rodolà
Neural networks embed the geometric structure of a data manifold lying in a high-dimensional space into latent representations. Ideally, the distribution of the data points in the latent space should depend only on the task, the data, the loss, and other architecture-specific constraints. However, factors such as the random weights initialization, training hyperparameters, or other sources of randomness in the training phase may induce incoherent latent spaces that hinder any form of reuse. Nevertheless, we empirically observe that, under the same data and modeling choices, the angles between the encodings within distinct latent spaces do not change. In this work, we propose the latent similarity between each sample and a fixed set of anchors as an alternative data representation, demonstrating that it can enforce the desired invariances without any additional training. We show how neural architectures can leverage these relative representations to guarantee, in practice, invariance to latent isometries and rescalings, effectively enabling latent space communication: from zero-shot model stitching to latent space comparison between diverse settings. We extensively validate the generalization capability of our approach on different datasets, spanning various modalities (images, text, graphs), tasks (e.g., classification, reconstruction) and architectures (e.g., CNNs, GCNs, transformers).

****

oViT: An Accurate Second-Order Pruning Framework for Vision Transformers
Denis Kuznedelev,Eldar Kurtic,Elias Frantar,Dan Alistarh
Models from the Vision Transformer (ViT) family have recently provided breakthrough results across image classification tasks such as ImageNet. Yet, they still face barriers to deployment, notably the fact that their accuracy can be severel

y impacted by compression techniques such as pruning. In this paper, we take a step towards addressing this issue by introducing \textit{Optimal ViT Surgeon (oViT)}, a new state-of-the-art method for the weight sparsification of Vision Transformers (ViT) models. At the technical level, oViT introduces a new weight pruning algorithm which leverages second-order information, specifically adapted to be both highly-accurate and efficient in the context of ViTs. We complement this accurate one-shot pruner with an in-depth investigation of gradual pruning, augmentation, and recovery schedules for ViTs, which we show to be critical for successful ViT compression. We validate our method via extensive experiments on classical ViT and DeiT models, as well as on newer variants, such as XCiT, EfficientFormer and Swin. Moreover, our results are even relevant to recently-proposed highly-accurate ResNets. Our results show for the first time that ViT-family models can in fact be pruned to high sparsity levels (e.g. $\geq 75\%$) with low impact on accuracy ($\leq 1\%$ relative drop), and that our approach outperforms prior methods by significant margins at high sparsities. In addition, we show that our method is compatible with structured pruning methods and quantization, and that it can lead to significant speedups on a sparsity-aware inference engine.

****

Learning Basic Interpretable Factors from Temporal Signals via Physics Symmetry
Xuanjie Liu,Daniel Chin,Yichen Huang,Gus Xia
We have recently seen great progress in learning interpretable music representations, ranging from basic factors, such as pitch and timbre, to high-level concepts, such as chord, texture and melody contour. However, most methods rely heavily on music domain knowledge and it remains an open question how to learn interpretable and disentangled representations using inductive biases that are more general. In this study, we use \textit{physical symmetry} as a self-consistency constraint on the latent space. Specifically, it requires the prior model that characterises the dynamics of the latent states to be \textit{equivariant} with respect to a certain group transformation (say, translation and rotation). We show that our model can learn \textit{linear} pitch factor (that agrees with human music perception) as well as pitch-timbre disentanglement from unlabelled monophonic music audio. In addition, the same methodology can be applied to computer vision, learning the 3D Cartesian space as well as space-colour disentanglement from a simple moving object shot by a single fix camera. Furthermore, applying physical symmetry to the prior model naturally leads to \textit{representation augmentation}, a new learning technique which helps improve sample efficiency.

****

Addressing High-dimensional Continuous Action Space via Decomposed Discrete Policy-Critic
Yechen Zhang,Jian Sun,Gang Wang,Jie Chen
Reinforcement learning (RL) methods for discrete action spaces like DQNs are being widely used in tasks such as Atari games. However, they encounter difficulties when addressing continuous control tasks, since discretizing continuous action space incurs the curse-of-dimensionality. To tackle continuous control tasks via discretized actions, we propose a decomposed discrete policy-critic (D2PC) architecture, which was inspired by multi-agent RL (MARL) and associates with each action dimension a discrete policy, while leveraging a single critic network to provide a shared evaluation. Building on D2PC, we advocate soft stochastic D2PC (SD2PC) and deterministic D2PC (D3PC) methods with a discrete stochastic or deterministic policy, which show comparable or superior training performances relative to even continuous actor-critic methods. Additionally, we design a mechanism that allows D3PC to interact with continuous actor-critic methods, contributing to the Q-policy-critic (QPC) algorithm, which inherits the training efficiency of discrete RL and the near-optimal final performance of continuous RL algorithms. Substantial experimental results on several continuous benchmark tasks validate our claims.

****

Unsupervised Manifold Alignment with Joint Multidimensional Scaling
Dexiong Chen,Bowen Fan,Carlos Oliver,Karsten Borgwardt

We introduce Joint Multidimensional Scaling, a novel approach for unsupervised manifold alignment, which maps datasets from two different domains, without any known correspondences between data instances across the datasets, to a common low-dimensional Euclidean space. Our approach integrates Multidimensional Scaling (MDS) and Wasserstein Procrustes analysis into a joint optimization problem to simultaneously generate isometric embeddings of data and learn correspondences between instances from two different datasets, while only requiring intra-dataset pairwise dissimilarities as input. This unique characteristic makes our approach applicable to datasets without access to the input features, such as solving the inexact graph matching problem. We propose an alternating optimization scheme to solve the problem that can fully benefit from the optimization techniques for MDS and Wasserstein Procrustes. We demonstrate the effectiveness of our approach in several applications, including joint visualization of two datasets, unsupervised heterogeneous domain adaptation, graph matching, and protein structure alignment. The implementation of our work is available at https://github.com/BorgwardtLab/JointMDS.

****

Can Single-Pass Contrastive Learning Work for Both Homophilic and Heterophilic Graph?

Haonan Wang,Jieyu Zhang,Qi Zhu,Wei Huang

Existing graph contrastive learning (GCL) typically requires two forward pass for a single instance to construct the contrastive loss. Despite its remarkable success, it is unclear whether such a dual-pass design is (theoretically) necessary. Besides, the empirical results are hitherto limited to the homophilic graph benchmarks. Then a natural question arises: Can we design a method that works for both homophilic and heterophilic graphs with a performance guarantee? To answer this, we theoretically analyze the concentration property of features obtained by neighborhood aggregation on both homophilic and heterophilic graphs, introduce the single-pass graph contrastive learning loss based on the property, and provide performance guarantees of the minimizer of the loss on downstream tasks. As a direct consequence of our theory, we introduce the Single-Pass Graph Contrastive Learning method (SP-GCL). Empirically, on 14 benchmark datasets with varying degrees of heterophily, the features learned by the SP-GCL can match or outperform existing strong baselines with significantly less computational overhead, and empirical results show the feasibility of conclusions derived by our analysis in real-world cases.

****

TOAST: Topological Algorithm for Singularity Tracking

Julius Von Rohrscheidt,Bastian Rieck

The manifold hypothesis, which assumes that data lie on or close to an unknown manifold of low intrinsic dimensionality, is a staple of modern machine learning research. However, recent work has shown that real-world data exhibit distinct non-manifold structures, which result in singularities that can lead to erroneous conclusions about the data. Detecting such singularities is therefore crucial as a precursor to interpolation and inference tasks. We address detecting singularities by developing (i) persistent local homology, a new topology-driven framework for quantifying the intrinsic dimension of a data set locally, and (ii) Euclidicity, a topology-based multi-scale measure for assessing the ‘manifoldness’ of individual points. We show that our approach can reliably identify singularities of complex spaces, while also capturing singular structures in real-world data sets.

****

Robust Manifold Estimation Approach for Evaluating Fidelity and Diversity

Pum Jun Kim,Yoojin Jang,Jisu Kim,Jaejun Yoo

We propose a robust and reliable evaluation metric for generative models by introducing topological and statistical treatments for a rigorous support manifold estimation. Existing metrics, such as Inception Score (IS), Frechet Inception Distance (FID), and the variants of Precision and Recall (P&R), heavily rely on support manifolds that are estimated from sample features. However, the reliability of their estimation has not been seriously discussed (and overlooked) even thou gh the quality of the evaluation entirely depends on it. In this paper, we propose Topological Precision and Recall (TopP&R, pronounced “topper”), which provides a systematic approach to estimating support manifolds, retaining only topologically and statistically important features with a certain level of confidence. This not only makes TopP&R strong for noisy features, but also provides statistical consistency. Our theoretical and experimental results show that TopP&R is robust to outliers and non-independent and identically distributed (Non-IID) perturbations, while accurately capturing the true trend of change in samples. To the best of our knowledge, this is the first evaluation metric focused on the robust estimation of the support manifold and provides its statistical consistency under noise.

****

Exploiting Certified Defences to Attack Randomised Smoothing

Andrew Craig Cullen,Paul Montague,Shijie Liu,Sarah Monazam Erfani,Benjamin I. P. Rubinstein

Certified guarantees of adversarial robustness play an important role in providing assurances regarding a models output, irrespective of the behaviour of an attacker. However, while the development of such guarantees has drawn upon an improved understanding of attacker behaviour, so too can certified guarantees be exploited in order to generate more efficient adversarial attacks. Within this work, we explore this heretofore undiscovered additional attack surface, while also considering how previously discovered attacks could be applied to models defended by randomised smoothing. In all bar one experiment our approach generates smaller adversarial perturbations for more than $70 \%$ of tested samples, reducing the average magnitude of the adversarial perturbation by $13 \%$.

****

Simple and Scalable Nearest Neighbor Machine Translation

Yuhan Dai,Zhirui Zhang,Qiuzhi Liu,Qu Cui,Weihua Li,Yichao Du,Tong Xu

$k$NN-MT is a straightforward yet powerful approach for fast domain adaptation, which directly plugs the pre-trained neural machine translation (NMT) models with domain-specific token-level $k$-nearest-neighbor ($k$NN) retrieval to achieve domain adaptation without retraining. Despite being conceptually attractive, $k$NN-MT is burdened with massive storage requirements and high computational complexity since it conducts nearest neighbor searches over the entire reference corpus. In this paper, we propose a simple and scalable nearest neighbor machine translation framework to drastically promote the decoding and storage efficiency of $k$NN-based models while maintaining the translation performance. To this end, we dynamically construct a extremely small datastore for each input via sentence-level retrieval to avoid searching the entire datastore in vanilla $k$NN-MT, based on which we further introduce a distance-aware adapter to adaptively incorporate the $k$NN retrieval results into the pre-trained NMT models. Experiments on machine translation in two general settings, static domain adaptation, and online learning, demonstrate that our proposed approach not only achieves almost 90% speed as the NMT model without performance degradation, but also significantly reduces the storage requirements of $k$NN-MT.

****

On the Effectiveness of Out-of-Distribution Data in Self-Supervised Long-Tail Learning.

Jianhong Bai,Zuozhu Liu,Hualiang Wang,Jin Hao,YANG FENG,Huanpeng Chu,Haoji Hu

Though Self-supervised learning (SSL) has been widely studied as a promising technique for representation learning, it doesn't generalize well on long-tailed datasets due to the majority classes dominating the feature space. Recent work shows that the long-tailed learning performance could be boosted by sampling extra in-domain (ID) data for self-supervised training, however, large-scale ID data which can rebalance the minority classes are expensive to collect. In this paper, we propose an alternative but easy-to-use and effective solution, \textbf{C}ontrastive with \textbf{O}ut-of-distribution (OOD) data for \textbf{L}ong-\textbf{T}ail learning (COLT), which can effectively exploit OOD data to dynamically re-balance the feature space. We empirically identify the counter-intuitive usefulness of OOD samples in SSL long-tailed learning and principally design a novel SSL

method. Concretely, we first localize the `\emph{head}' and `\emph{tail}' samples by assigning a tailness score to each OOD sample based on its neighborhoods in the feature space. Then, we propose an online OOD sampling strategy to dynamically re-balance the feature space. Finally, we enforce the model to be capable of distinguishing ID and OOD samples by a distribution-level supervised contrastive loss. Extensive experiments are conducted on various datasets and several state-of-the-art SSL frameworks to verify the effectiveness of the proposed method. The results show that our method significantly improves the performance of SSL on long-tailed datasets by a large margin, and even outperforms previous work which uses external ID data. Our code is available at \url{https://github.com/JianhongBai/COLT}.
****
Dynamic Update-to-Data Ratio: Minimizing World Model Overfitting
Nicolai Dorka,Tim Welschehold,Wolfram Burgard
Early stopping based on the validation set performance is a popular approach to find the right balance between under- and overfitting in the context of supervised learning. However, in reinforcement learning, even for supervised sub-problems such as world model learning, early stopping is not applicable as the dataset is continually evolving. As a solution, we propose a new general method that dynamically adjusts the update to data (UTD) ratio during training based on under- and overfitting detection on a small subset of the continuously collected experience not used for training. We apply our method to DreamerV2, a state-of-the-art model-based reinforcement learning algorithm, and evaluate it on the DeepMind Control Suite and the Atari 100k benchmark. The results demonstrate that one can better balance under- and overestimation by adjusting the UTD ratio with our approach compared to the default setting in DreamerV2 and that it is competitive with an extensive hyperparameter search which is not feasible for many applications. Our method eliminates the need to set the UTD hyperparameter by hand and even leads to a higher robustness with regard to other learning-related hyperparameters further reducing the amount of necessary tuning.
****
Uni-Mol: A Universal 3D Molecular Representation Learning Framework
Gengmo Zhou,Zhifeng Gao,Qiankun Ding,Hang Zheng,Hongteng Xu,Zhewei Wei,Linfeng Zhang,Guolin Ke
Molecular representation learning (MRL) has gained tremendous attention due to its critical role in learning from limited supervised data for applications like drug design. In most MRL methods, molecules are treated as 1D sequential tokens or 2D topology graphs, limiting their ability to incorporate 3D information for downstream tasks and, in particular, making it almost impossible for 3D geometry prediction/generation. In this paper, we propose a universal 3D MRL framework, called Uni-Mol, that significantly enlarges the representation ability and application scope of MRL schemes. Uni-Mol contains two pretrained models with the same SE(3) Transformer architecture: a molecular model pretrained by 209M molecular conformations; a pocket model pretrained by 3M candidate protein pocket data. Besides, Uni-Mol contains several finetuning strategies to apply the pretrained models to various downstream tasks. By properly incorporating 3D information, Uni-Mol outperforms SOTA in 14/15 molecular property prediction tasks. Moreover, Uni-Mol achieves superior performance in 3D spatial tasks, including protein-ligand binding pose prediction, molecular conformation generation, etc. The code, model, and data are made publicly available at https://github.com/dptech-corp/Uni-Mol.
****
CAPE: Channel-Attention-Based PDE Parameter Embeddings for SciML
Makoto Takamoto,Francesco Alesiani,Mathias Niepert
Scientific Machine Learning (SciML) designs machine learning methods that predict physical systems governed by partial differential equations (PDE). These ML-based surrogate models substitute inefficient and often non-differentiable numerical simulation algorithms and find multiple applications such as weather forecasting, molecular dynamics, and medical applications.
While a number of ML-based methods for approximating the solutions of PDEs have

been proposed in recent years, they typically do not consider the parameters of the PDEs, making it difficult for the ML surrogate models to generalize to PDE parameters not seen during training.

We propose a new channel-attention-based parameter embedding (CAPE) component for scientific machine learning models and a simple and effective curriculum learning strategy. The CAPE module can be combined with any kind of ML surrogate model, which can adapt to changing PDE parameters without harming the original model's ability to find approximate solutions to PDEs. The curriculum learning strategy provides a seamless transition between teacher-forcing and fully auto-regressive training.

We compare CAPE in conjunction with the curriculum learning strategy using a PDE benchmark and obtain consistent and significant improvements over the base models. The experiments also show several advantages of CAPE, such as its increased ability to generalize to unseen PDE parameters without substantially increasing inference time and parameter count.

An implementation of the method and experiments are available at \url{https://anonymous.4open.science/r/CAPE-ML4Sci-145B}.

****

Topic and Hyperbolic Transformer to Handle Multi-modal Dependencies

Noriaki Kawamae

As multi-modal search relies on jointly learning image-text representations and has been investigated in the literature,

our innovation is to develop Chimera, a framework in which to learn their representations and similarities.

Because the core of multi-modal search is learning the modalities in a shared semantic space and measuring their similarities,

search quality depends on which expressive space is utilized in learning.

This motivates us to identify the space that can elucidate their semantic and complex relationships with small information loss.

Novelty is assured by introducing the topic and hyperbolic as spaces,

and performing contrastive/metric learning tasks to ensure the cooperation of these spaces with Transformer.

Experiments show that Chimera empowers pre-trained models for multi-modal search tasks and demonstrate the ability of the layers it introduces.

****

Variance Covariance Regularization Enforces Pairwise Independence in Self-Supervised Representations

Grégoire Mialon,Randall Balestriero,Yann LeCun

Self-Supervised Learning (SSL) methods such as VICReg, Barlow Twins or W-MSE avoid collapse of their joint embedding architectures by constraining or regularizing the covariance matrix of their projector’s output. This study highlights important properties of such strategy, which we coin Variance-Covariance regularization (VCReg). More precisely, we show that VCReg enforces pairwise independence between the features of the learned representation. This result emerges by bridging VCReg applied on the projector’s output to kernel independence criteria applied on the projector’s input. This provides the first theoretical motivations and explanations of VCReg. We empirically validate our findings where (i) we put in evidence which projector’s characteristics favor pairwise independence, (ii) we use these findings to obtain nontrivial performance gains for VICReg, (iii) we demonstrate that the scope of VCReg goes beyond SSL by using it to solve Independent Component Analysis. We hope that our findings will support the adoption of VCReg in SSL and beyond.

****

DEP-RL: Embodied Exploration for Reinforcement Learning in Overactuated and Musculoskeletal Systems

Pierre Schumacher,Daniel Haeufle,Dieter Büchler,Syn Schmitt,Georg Martius

Muscle-actuated organisms are capable of learning an unparalleled diversity of dexterous movements despite their vast amount of muscles.

Reinforcement learning (RL) on large musculoskeletal models, however, has not be

en able to show similar performance.
We conjecture that ineffective exploration in large overactuated action spaces is a key problem.
This is supported by the finding that common exploration noise strategies are inadequate in synthetic examples of overactuated systems.
We identify differential extrinsic plasticity (DEP), a method from the domain of self-organization, as being able to induce state-space covering exploration within seconds of interaction.
By integrating DEP into RL, we achieve fast learning of reaching and locomotion in musculoskeletal systems, outperforming current approaches in all considered tasks in sample efficiency and robustness.
****

Restricted Generative Projection for One-Class Classification and Anomaly detection
Feng Xiao,Ruoyu Sun,Jicong Fan
We present a novel framework for one-class classification and anomaly detection. The core idea is to learn a mapping to transform the unknown distribution of training (normal) data to a known distribution that is supposed to be different from the transformed distribution of unknown abnormal data. Crucially, the target distribution of training data should be sufficiently simple, compact, and informative. The simplicity is to ensure that we can sample from the distribution easily, the compactness is to ensure that the decision boundary between normal data and abnormal data is clear and reliable, and the informativeness is to ensure that the transformed data preserve the important information of the original data. Therefore, we propose to use truncated Gaussian, uniform in hyperball, uniform on hypersphere, or uniform between hyperspheres, as the target distribution. We then minimize the distance between the transformed data distribution and the target distribution while keeping the reconstruction error for the original data small enough. Our model is simple and easy to train especially compared with those based on generative models. Comparative studies on a few benchmark datasets verify the effectiveness of our method in comparison to baselines.
****

Name Your Colour For the Task: Artificially Discover Colour Naming via Colour Quantisation Transformer
Shenghan Su,Lin Gu,Ziteng Cui,Yue Yang,Jingjing Shen,Hiroaki Yamane,Zenghui Zhang,Tatsuya Harada
The long-standing theory that a colour-naming system evolves under the dual pressure of efficient communication and perceptual mechanism is supported by more and more linguistic studies including the analysis of four decades’ diachronic data from the Nafaanra language. This inspires us to explore whether artificial intelligence could evolve and discover a similar colour-naming system via optimising the communication efficiency represented by high-level recognition performance. Here, we propose a novel colour quantisation transformer, CQFormer, that quantises colour space while maintaining the accuracy of machine recognition on the quantised image. Given an RGB image, the annotation branch maps it into an index map before generating the quantised image with a colour palette, meanwhile the palette branch utilises a key-point detection way to find proper colours in palette among whole colour space. By interacting with colour annotation, CQFormer is able to balance both the machine vision accuracy and colour perceptual structure such as distinct and stable colour distribution for discovered colour system. Very interestingly, we even observe the consistent evolution pattern between our artificial colour system and basic colour terms across human languages. Besides, our colour quantisation method also offers an efficient quantisation method that effectively compresses the image storage while maintaining a high performance in high-level recognition tasks such as classification and detection. Extensive experiments demonstrate the superior performance of our method with extremely low bit-rate colours. We will release the source code upon acceptance.
****

The Symmetric Generalized Eigenvalue Problem as a Nash Equilibrium
Ian Gemp,Charlie Chen,Brian McWilliams

The symmetric generalized eigenvalue problem (SGEP) is a fundamental concept in numerical linear algebra. It captures the solution of many classical machine learning problems such as canonical correlation analysis, independent components analysis, partial least squares, linear discriminant analysis, principal components and others. Despite this, most general solvers are prohibitively expensive when dealing with *streaming data sets* (i.e., minibatches) and research has instead concentrated on finding efficient solutions to specific problem instances. In this work, we develop a game-theoretic formulation of the top-k SGEP whose Nash equilibrium is the set of generalized eigenvectors. We also present a parallelizable algorithm with guaranteed asymptotic convergence to the Nash. Current state-of-the-art methods require $\mathcal{O}(d^2k)$ runtime complexity per iteration which is prohibitively expensive when the number of dimensions (d) is large. We show how to modify this parallel approach to achieve $\mathcal{O}(dk)$ runtime complexity. Empirically we demonstrate that this resulting algorithm is able to solve a variety of SGEP problem instances including a large-scale analysis of neural network activations.

****

Learning with Auxiliary Activation for Memory-Efficient Training

Sunghyeon Woo,Dongsuk Jeon

While deep learning has achieved great success in various fields, a large amount of memory is necessary to train deep neural networks, which hinders the development of massive state-of-the-art models. The reason is the conventional learning rule, backpropagation, should temporarily store input activations of all the layers in the network. To overcome this, recent studies suggested various memory-efficient implementations of backpropagation. However, those approaches incur computational overhead due to the recomputation of activations, slowing down neural network training. In this work, we propose a new learning rule which significantly reduces memory requirements while closely matching the performance of backpropagation. The algorithm combines auxiliary activation with output activation during forward propagation, while only auxiliary activation is used during backward propagation instead of actual input activation to reduce the amount of data to be temporarily stored. We mathematically show that our learning rule can reliably train the networks whose loss landscape is convex if the auxiliary activation satisfies certain conditions. Based on this observation, we suggest candidates of auxiliary activation that satisfy those conditions. Experimental results confirm that the proposed learning rule achieves competitive performance compared to backpropagation in various models such as ResNet, Transformer, BERT, ViT, and MLP-Mixer.

****

Equivariant 3D-Conditional Diffusion Models for Molecular Linker Design

Ilia Igashov,Hannes Stärk,Clement Vignac,Victor Garcia Satorras,Pascal Frossard,Max Welling,Michael M. Bronstein,Bruno Correia

Fragment-based drug discovery has been an effective paradigm in early-stage drug development. An open challenge in this area is designing linkers between disconnected molecular fragments of interest to obtain chemically-relevant candidate drug molecules. In this work, we propose DiffLinker, an E(3)-equivariant 3D-conditional diffusion model for molecular linker design. Given a set of disconnected fragments, our model places missing atoms in between and designs a molecule incorporating all the initial fragments. Unlike previous approaches that are only able to connect pairs of molecular fragments, our method can link an arbitrary number of fragments. Additionally, the model automatically determines the number of atoms in the linker and its attachment points to the input fragments. We demonstrate that DiffLinker outperforms other methods on the standard datasets generating more diverse and synthetically-accessible molecules. Besides, we experimentally test our method in real-world applications, showing that it can successfully generate valid linkers conditioned on target protein pockets.

****

Language Modelling with Pixels

Phillip Rust,Jonas F. Lotz,Emanuele Bugliarello,Elizabeth Salesky,Miryam de Lhoneux,Desmond Elliott

Language models are defined over a finite set of inputs, which creates a vocabulary bottleneck when we attempt to scale the number of supported languages. Tackling this bottleneck results in a trade-off between what can be represented in the embedding matrix and computational issues in the output layer. This paper introduces PIXEL, the Pixel-based Encoder of Language, which suffers from neither of these issues. PIXEL is a pretrained language model that renders text as images, making it possible to transfer representations across languages based on orthographic similarity or the co-activation of pixels. PIXEL is trained to reconstruct the pixels of masked patches instead of predicting a distribution over tokens. We pretrain the 86M parameter PIXEL model on the same English data as BERT and evaluate on syntactic and semantic tasks in typologically diverse languages, including various non-Latin scripts. We find that PIXEL substantially outperforms BERT on syntactic and semantic processing tasks on scripts that are not found in the pretraining data, but PIXEL is slightly weaker than BERT when working with Latin scripts. Furthermore, we find that PIXEL is more robust than BERT to orthographic attacks and linguistic code-switching, further confirming the benefits of modelling language with pixels.
****

Sinkhorn Discrepancy for Counterfactual Generalization
Hao Wang,Quanyu Dai,Jiajun Fan,Weiming Liu,Zhichao Chen,Tianqiao Liu,Yichao Wang,Zhenhua Dong,Ruiming Tang

Estimating individual treatment effects from observational data is very challenging due to the existence of treatment selection bias.

Most existing representation-based methods mitigate this issue by aligning distributions of different treatment groups in the representation space. However, they still suffer from two critical problems: (1) Mini-batch Sampling Effects (MSE), where the alignment easily fails due to the outcome imbalance or outliers in the batch; (2) Unobserved Confounder Effects (UCE), where the unobserved confounders damage the correct alignment. To tackle these problems, we propose a principled approach named Entire Space CounterFactual Regression (ESCFR) based on a generalized sinkhorn discrepancy for distribution alignment within the stochastic optimal transport framework. Based on the framework, we propose a relaxed mass preserving regularizer to address the MSE issue and design a proximal factual outcome regularizer to handle the UCE issue. Extensive experiments demonstrate that our proposed ESCFR can successfully tackle the treatment selection bias and achieve significantly better performance than state-of-the-art methods.
****

Massively Scaling Heteroscedastic Classifiers
Mark Collier,Rodolphe Jenatton,Basil Mustafa,Neil Houlsby,Jesse Berent,Effrosyni Kokiopoulou

Heteroscedastic classifiers, which learn a multivariate Gaussian distribution over prediction logits, have been shown to perform well on image classification problems with hundreds to thousands of classes. However, compared to standard classifiers, they introduce extra parameters that scale linearly with the number of classes. This makes them infeasible to apply to larger-scale problems. In addition heteroscedastic classifiers introduce a critical temperature hyperparameter which must be tuned. We propose HET-XL, a heteroscedastic classifier whose parameter count when compared to a standard classifier scales independently of the number of classes. In our large-scale settings, we show that we can remove the need to tune the temperature hyperparameter, by directly learning it on the training data. On large image classification datasets with up to 4B images and 30k classes our method requires 14X fewer additional parameters, does not require tuning the temperature on a held-out set and performs consistently better than the baseline heteroscedastic classifier. HET-XL improves ImageNet 0-shot classification in a multimodal contrastive learning setup which can be viewed as a 3.5 billion class classification problem.
****

Vera Verto: Multimodal Hijacking Attack
Minxing Zhang,Ahmed Salem,Michael Backes,Yang Zhang

The increasing cost of training machine learning (ML) models has led to the incl

usion of new parties to the training pipeline, such as users who contribute training data and companies that provide computing resources. This involvement of such new parties in the ML training process has introduced new attack surfaces for an adversary to exploit. A recent attack in this domain is the model hijacking attack, whereby an adversary hijacks a victim model to implement their own -- possibly malicious -- hijacking tasks. However, the scope of the model hijacking attack is so far limited to computer vision-related tasks. In this paper, we transform the model hijacking attack into a more general multimodal setting, where the hijacking and original tasks are performed on data of different modalities. Specifically, we focus on the setting where an adversary implements a natural language processing (NLP) hijacking task into an image classification model. To mount the attack, we propose a novel encoder-decoder based framework, namely the Blender, which relies on advanced image and language models. Experimental results show that our modal hijacking attack achieves strong performances in different settings. For instance, our attack achieves 94%, 94%, and 95% attack success rate when using the Sogou news dataset to hijack STL10, CIFAR-10, and MNIST classifiers.
****

Joint Attention-Driven Domain Fusion and Noise-Tolerant Learning for Multi-Source Domain Adaptation

Tong Xu,Lin Wang,Wu Ning,Chunyan Lyu,Kejun Wang

Multi-source Unsupervised Domain Adaptation (MUDA) transfers knowledge from multiple source domains with labeled data to an unlabeled target domain.

Recently, endeavours have been made in establishing connections among different domains to enable feature interaction. However, these approaches essentially enhance category information and thus lack the transfer of the domain-specific information. Moreover, few research has explored the connection between pseudo-label generation and the framework’s learning capabilities, crucial for ensuring robust MUDA. In this paper, we propose a novel framework, which significantly reduces the domain discrepancy and demonstrates new state-of-the-art performance. In particular, we first propose a Contrary Attention-based Domain Merge (CADM) module to enable the interaction among the features so as to achieve the mixture of domain-specific information instead of focusing on the category information. Secondly, to enable the network to correct the pseudo labels during training, we propose an adaptive and reverse cross-entropy loss, which can adaptively impose constraints on the pseudo-label generation process. We conduct experiments on four benchmark datasets, showing that our approach can efficiently fuse all domains for MUDA while showing much better performance than the prior methods.
****

EA-HAS-Bench: Energy-aware Hyperparameter and Architecture Search Benchmark

Shuguang Dou,XINYANG JIANG,Cai Rong Zhao,Dongsheng Li

The energy consumption for training deep learning models is increasing at an alarming rate due to the growth of training data and model scale, resulting in a negative impact on carbon neutrality. Energy consumption is an especially pressing issue for AutoML algorithms because it usually requires repeatedly training large numbers of computationally intensive deep models to search for optimal configurations. This paper takes one of the most essential steps in developing energy-aware (EA) NAS methods, by providing a benchmark that makes EA-NAS research more reproducible and accessible. Specifically, we present the first large-scale energy-aware benchmark that allows studying AutoML methods to achieve better trade-offs between performance and search energy consumption, named EA-HAS-Bench. EA-HAS-Bench provides a large-scale architecture/hyperparameter joint search space, covering diversified configurations related to energy consumption. Furthermore, we propose a novel surrogate model specially designed for large joint search space, which proposes a Bezier curve-based model to predict learning curves with unlimited shape and length. Based on the proposed dataset, we new energy-aware AutoML method that arms existing AutoML algorithms to consider the search energy consumption, and our experiments show that the modified energy-aware AutoML methods achieve a better trade-off between energy consumption and model performance.
****

Breaking the Curse of Dimensionality for Parametric Elliptic PDEs
Marius Zeinhofer,Alex Kaltenbach

Motivated by recent empirical success, we examine how neural network-based ansatz classes can break the curse of dimensionality for high-dimensional, non-linear elliptic partial differential equations (PDEs) with variational structure. The high-dimensionality of the PDEs can either be induced through a high-dimensional physical domain or a high-dimensional parameter space. The latter include parametric right-hand sides, parametric domains, and material constants. Our main result shows that any scheme, that computes neural network based $W^{1,p}$-approximations, leverages the extraordinary approximation capabilities of neural networks and, thus, is able to beat the curse of dimensionality if the ground truth solution is smooth or possesses Barron regularity. Popular examples of $W^{1,p}$-convergent schemes include, e.g., the Deep Ritz Method and physics-informed neural networks. We present numerical experiments supporting our theoretical findings.

****

UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer
Kunchang Li,Yali Wang,Yinan He,Yizhuo Li,Yi Wang,Limin Wang,Yu Qiao

Learning discriminative spatiotemporal representation is the key problem of video understanding. Recently, Vision Transformers (ViTs) have shown their power in learning long-term video dependency with self-attention. Unfortunately, they exhibit limitations in tackling local video redundancy, due to the blind global comparison among tokens. UniFormer has successfully alleviated this issue, by unifying convolution and self-attention as a relation aggregator in the transformer format. However, this model has to require a tiresome and complicated image-pretraining phrase, before being finetuned on videos. This blocks its wide usage in practice. On the contrary, open-sourced ViTs are readily available and well-pretrained with rich image supervision. Based on these observations, we propose a generic paradigm to build a powerful family of video networks, by arming the pretrained ViTs with efficient UniFormer designs. We call this family UniFormerV2, since it inherits the concise style of the UniFormer block. But it contains brand-new local and global relation aggregators, which allow for preferable accuracy-computation balance by seamlessly integrating advantages from both ViTs and UniFormer. Without any bells and whistles, our UniFormerV2 gets the state-of-the-art recognition performance on 8 popular video benchmarks, including scene-related Kinetics-400/600/700 and Moments in Time, temporal-related Something-Something V1/V2, untrimmed ActivityNet and HACS. In particular, it is the first model to achieve 90% top-1 accuracy on Kinetics-400, to our best knowledge. The models will be released afterward.

****

Dynamical Equations With Bottom-up Self-Organizing Properties Learn Accurate Dynamical Hierarchies Without Any Loss Function
Danilo Vasconcellos Vargas,Tham Yik Foong,Heng Zhang

Self-organization is ubiquitous in nature and mind. However, machine learning and theories of cognition still barely touch the subject. The hurdle is that general patterns are difficult to define in terms of dynamical equations and designing a system that could learn by reordering itself is still to be seen. Here, we propose a learning system, where patterns are defined within the realm of nonlinear dynamics with positive and negative feedback loops, allowing attractor-repeller pairs to emerge for each pattern observed. Experiments reveal that such a system can map temporal to spatial correlation, enabling hierarchical structures to be learned from sequential data. The results are accurate enough to surpass state-of-the-art unsupervised learning algorithms in seven out of eight experiments as well as two real-world problems. Interestingly, the dynamic nature of the system makes it inherently adaptive, giving rise to phenomena similar to phase transitions in chemistry/thermodynamics when the input structure changes. Thus, the work here sheds light on how self-organization can allow for pattern recognition and hints at how intelligent behavior might emerge from simple dynamic equations without an objective/loss function.

****

Multi-Label Knowledge Distillation

Peng-Hui Yang,Ming-Kun Xie,Chen-Chen Zong,Lei Feng,Gang Niu,Masashi Sugiyama,Sheng-Jun Huang

Existing knowledge distillation methods typically work by enforcing the consistency of output logits or intermediate feature maps between the teacher network and student network. Unfortunately, these methods can hardly be extended to the multi-label learning scenario. Because each instance is associated with multiple semantic labels, neither the prediction logits nor the feature maps obtained from the whole example can accurately transfer knowledge for each label. In this paper, we propose a novel multi-label knowledge distillation method. On one hand, it exploits the informative semantic knowledge from the logits by label decoupling with the one-versus-all reduction strategy; on the other hand, it enhances the distinctiveness of the learned feature representations by leveraging the structural information of label-wise embeddings. Experimental results on multiple benchmark datasets validate that the proposed method can avoid knowledge counteraction among labels, and achieve superior performance against diverse comparing methods.

****

ADVERSARY-AWARE PARTIAL LABEL LEARNING WITH LABEL DISTILLATION

Cheng Chen,Yueming Lyu,Ivor Tsang

To ensure that the data collected from human subjects is entrusted with a secret, rival labels are introduced to conceal the information provided by the participants on purpose. The corresponding learning task can be formulated as a noisy partial-label learning problem. However, conventional partial-label learning (PLL) methods are still vulnerable to the high ratio of noisy partial labels, especially in a large labelling space. To learn a more robust model, we present Adversary-Aware Partial Label Learning and introduce the $\textit{rival}$, a set of noisy labels, to the collection of candidate labels for each instance. By introducing the rival label, the predictive distribution of PLL is factorised such that a reasonably good predictive label is achieved with less uncertainty coming from the transition matrix, assuming its generation process is known. Nonetheless, the predictive accuracy is still insufficient to produce an adequately good set of positive samples to minimise the loss function. Moreover, the inclusion of rivals also brings an inconsistency issue for the classifier and risk function due to the intractability of the transition matrix. Consequently, the immature teacher within momentum (ITWM) disambiguation algorithm is proposed to cope with the situation. We utilise the confidence score mapping from the instance space to approximate the intractable term, allowing us to obtain a provably consistent classifier and risk function. Extensive experiments demonstrate that our method achieves promising results on the CIFAR10, CIFAR100 and CUB200 datasets.

****

Structural Privacy in Graphs

Rucha Bhalchandra Joshi,Subhankar Mishra

Graph Neural Networks (GNNs) gained popularity to address the tasks over the graph-structured data that best represent many real-world systems. The privacy of the participants of these systems is at risk if the GNNs are not carefully designed. Existing works in privacy-preserving GNNs primarily ensure the privacy of features and labels of a node. In order to ensure complete privacy related to graph data, its structure also needs to be privatized. We provide a method SPGraph to privatize the graph structure by adding noise to the neighborhood data of the node. Our method addresses two challenges in introducing structural privacy in graphs. Applying randomization on the set of actual neighbors to introduce noise leads to a reduction in the degree of a node, which is undesirable. To overcome this first challenge, we introduce λ-selector that samples nodes to be added to the set of neighbors. The second challenge is to denoise the neighborhood so that the noise added in the neighborhood does not significantly impact the accuracy. In this view, we use p-hop neighborhood to compensate for the loss of actual neighbors in the randomization. We continue to use the node and label privacy as implemented in the previous methods for privacy in GNNs. We conduct extensive experiments over real-world datasets to show the impact of perturbation in the graph structure.

****

KnowDA: All-in-One Knowledge Mixture Model for Data Augmentation in Low-Resource NLP

Yufei Wang,Jiayi Zheng,Can Xu,Xiubo Geng,Tao Shen,Chongyang Tao,Daxin Jiang

This paper focuses on data augmentation for low-resource NLP tasks where the training set is limited. The existing solutions either leverage task-independent heuristic rules (e.g., Synonym Replacement) or fine-tune general-purpose pre-trained language models (e.g., GPT2) using the limited training instances to produce new synthetic data. Consequently, they have trivial task-specific knowledge and are limited to yielding low-quality synthetic data. To combat this issue, we propose Knowledge Mixture Data Augmentation Model (KnowDA), a Seq2Seq language model pretrained on a mixture of diverse NLP tasks under a novel framework of Knowledge Mixture Training (KoMT). The goal of KoMT is to condense diverse NLP task-specific knowledge into the single KnowDA model (i.e., all-in-one). The resulting KnowDA could utilize these knowledge to quickly grasp the inherent synthesis law of the target task through limited training instances. Specifically, KoMT reformulates input examples from various heterogeneous NLP tasks into a unified text-to-text format and employs denoising training objectives in different granularity to learn to reconstruct partial or complete samples. To the best of our knowledge, we are the first to attempt to apply 100+ NLP multi-task training for data augmentation. Extensive experiments show that i) the synthetic data produced by KnowDA successfully improves the performance of the strong pre-trained language models (i.e., Bert, ALBert and Deberta) by a large margin on the low-resource NLP benchmark FewGLUE, CoNLL’03 and WikiAnn; ii) KnowDA successful transfer the task knowledge to NLP tasks whose types are seen and unseen in KoMT.

****

Learning Graph Neural Network Topologies

Avishkar Saha,Oscar Mendez Maldonado,Chris Russell,Richard Bowden

Graph convolutional networks (GCNs) enable end-to-end learning on graph structured data. However, many works begin by assuming a given graph structure. As the ideal graph structure is often unknown, this limits applicability. To address this, we present a novel end-to-end differentiable graph-generator which builds the graph topology on the fly. Our module can be readily integrated into existing pipelines involving graph convolution operations, replacing the predetermined or existing adjacency matrix with one that is learned, and optimised, as part of the general objective. As such it is applicable to any GCN. We show that integrating our module into both node classification and trajectory prediction pipelines improves accuracy across a range of datasets and backbones.

****

Finding the Global Semantic Representation in GAN through Fréchet Mean

Jaewoong Choi,Geonho Hwang,Hyunsoo Cho,Myungjoo Kang

The ideally disentangled latent space in GAN involves the global representation of latent space using semantic attribute coordinates. In other words, in this disentangled space, there exists the global semantic basis as a vector space where each basis component describes one attribute of generated images. In this paper, we propose an unsupervised method for finding this global semantic basis in the intermediate latent space in GANs. This semantic basis represents sample-independent meaningful perturbations that change the same semantic attribute of an image on the entire latent space. The proposed global basis, called Fréchet basis, is derived by introducing Fréchet mean to the local semantic perturbations in a latent space. Fréchet basis is discovered in two stages. First, the global semantic subspace is discovered by the Fréchet mean in the Grassmannian manifold of the local semantic subspaces. Second, Fréchet basis is found by optimizing a basis of the semantic subspace via the Fréchet mean in the Special Orthogonal Group. Experimental results demonstrate that Fréchet basis provides better semantic factorization and robustness compared to the previous methods. Moreover, we suggest the basis refinement scheme for the previous methods. The quantitative experiments show that the refined basis achieves better semantic factorization while constrained on the same semantic subspace given by the previous method.

****

Identical Initialization: A Universal Approach to Fast and Stable Training of Neural Networks

Yu Pan,Zekai Wu,Chaozheng Wang,Qifan Wang,Min Zhang,Zenglin Xu

A well-conditioned initialization is beneficial for training deep neural networks. However, existing initialization approaches do not simultaneously show robustness and universality. Specifically, even though the widely-used Xavier and Kaiming initialization approaches can generally fit a variety of networks, they fail to train residual networks without Batch Normalization for calculating an inappropriate scale on data-flow. On the other hand, some literature design stable initialization (e.g., Fixup and ReZero) based on dynamical isometry, an efficient learning mechanism. Nonetheless, these methods are specifically designed for either a non-residual structure or a residual block only, and even include extra auxiliary components, limiting their applicable range. Intriguingly, we find that the identity matrix is a feasible and universal solution to the aforementioned problems, as it adheres to dynamical isometry while remaining applicable to a wide range of models. Motivated by this, we develop Identical Initialization (IDInit), a sufficiently robust, universal, and fast-converging approach on the identity matrix. Empirical results on a variety of benchmarks show that IDInit is universal to various network types, and practically useful with good performance and fast convergence.

****

Addressing Parameter Choice Issues in Unsupervised Domain Adaptation by Aggregation

Marius-Constantin Dinu,Markus Holzleitner,Maximilian Beck,Hoan Duc Nguyen,Andrea Huber,Hamid Eghbal-zadeh,Bernhard A. Moser,Sergei Pereverzyev,Sepp Hochreiter,Werner Zellinger

We study the problem of choosing algorithm hyper-parameters in unsupervised domain adaptation, i.e., with labeled data in a source domain and unlabeled data in a target domain, drawn from a different input distribution. We follow the strategy to compute several models using different hyper-parameters, and, to subsequently compute a linear aggregation of the models. While several heuristics exist that follow this strategy, methods are still missing that rely on thorough theories for bounding the target error. In this turn, we propose a method that extends weighted least squares to vector-valued functions, e.g., deep neural networks. We show that the target error of the proposed algorithm is asymptotically not worse than twice the error of the unknown optimal aggregation. We also perform a large scale empirical comparative study on several datasets, including text, images, electroencephalogram, body sensor signals and signals from mobile phones. Our method outperforms deep embedded validation (DEV) and importance weighted validation (IWV) on all datasets, setting a new state-of-the-art performance for solving parameter choice issues in unsupervised domain adaptation with theoretical error guarantees. We further study several competitive heuristics, all outperforming IWV and DEV on at least five datasets. However, our method outperforms each heuristic on at least five of seven datasets.

****

MARS: Meta-learning as Score Matching in the Function Space

Krunoslav Lehman Pavasovic,Jonas Rothfuss,Andreas Krause

Meta-learning aims to extract useful inductive biases from a set of related datasets. In Bayesian meta-learning, this is typically achieved by constructing a prior distribution over neural network parameters. However, specifying families of computationally viable prior distributions over the high-dimensional neural network parameters is difficult. As a result, existing approaches resort to meta-learning restrictive diagonal Gaussian priors, severely limiting their expressiveness and performance. To circumvent these issues, we approach meta-learning through the lens of functional Bayesian neural network inference which views the prior as a stochastic process and performs inference in the function space. Specifically, we view the meta-training tasks as samples from the data-generating process and formalize meta-learning as empirically estimating the law of this stochastic process. Our approach can seamlessly acquire and represent complex prior know ledge by meta-learning the score function of the data-generating process marginals instead of parameter space priors. In a comprehensive benchmark, we demonstrate that our method achieves state-of-the-art performance in terms of predictive accuracy and substantial improvements in the quality of uncertainty estimates.

****

Faster Gradient-Free Methods for Escaping Saddle Points

Hualin Zhang,Bin Gu

Escaping from saddle points has become an important research topic in non-convex optimization. In this paper, we study the case when calculations of explicit gradients are expensive or even infeasible, and only function values are accessible.

Currently, there have two types of gradient-free (zeroth-order) methods based on random perturbation and negative curvature finding proposed to escape saddle points efficiently and converge to an ϵ-approximate second-order stationary point.

Nesterov's accelerated gradient descent (AGD) method can escape saddle points faster than gradient descent (GD) which have been verified in first-order algorithms. However, whether AGD could accelerate the gradient-free methods is still unstudied. To unfold this mystery, in this paper, we propose two accelerated variants for the two types of gradient-free methods of escaping saddle points. We show that our algorithms can find an ϵ-approximate second-order stationary point with $\tilde{\mathcal{O}}(1/\epsilon^{1.75})$ iteration complexity and $\tilde{\mathcal{O}}(d/\epsilon^{1.75})$ oracle complexity, where d is the problem dimension. Thus, our methods achieve a comparable convergence rate to their first-order counterparts and have fewer oracle complexity compared to prior derivative-free methods for finding second-order stationary points.

****

Symmetrical SyncMap for Imbalanced General Chunking Problems

Heng Zhang,Danilo Vasconcellos Vargas

Recently, SyncMap (2021) pioneered an approach to learn complex structures from sequences as well as adapt to any changes in underlying structures. Such approach, inspired by neuron group behaviors, is achieved by using self-organizing dynamical equations without any loss functions. Here we propose Symmetrical SyncMap that goes beyond the original work to show how to create dynamical equations and attractor-repeller points which are stable over the long run, even dealing with imbalanced continual general chunking problems (CGCPs). The main idea is to apply equal updates from positive and negative feedback loops by symmetrical activation. We then introduce the concept of memory window to allow for more positive updates. Our algorithm surpasses or ties other unsupervised state-of-the-art baselines in all 12 imbalanced CGCPs with various difficulties, including dynamical ones. To verify its performance in real-world scenarios, we conduct experiments on several well-studied structure learning problems. The proposed method surpasses substantially other methods in all scenarios, suggesting that symmetrical activation plays a critical role in uncovering topological structures and even hierarchies encoded in temporal data.

****

Solving Partial Label Learning Problem with Multi-Agent Reinforcement Learning

Xinyi Zhang,Xingdong Feng,Fan Zhou

Partial label learning (PLL) deals with classifications when a set of candidate labels instead of the true one is given for each training instance. As a weakly supervised learning problem, the main target of PLL is to discover latent relationships within training samples, and utilize such information to disambiguate noisy labels. Many existing methods choose nearest neighbors of each partially-labeled instance in an unsupervised way such that the obtained instance similarities can be empirically non-optimal and unrelated to the downstream classification task. To address this issue, we propose a novel multi-agent reinforcement learning (MARL) framework which models the connection between each pair of training samples as a reinforcement learning (RL) agent. We use attention-based graph neural network (GNN) to learn the instance similarity, and adaptively refine it using a deterministic policy gradient approach until some pre-defined scoring functio n is optimized. Different from those two-stage and alternative optimization algorithms whose training procedures are not end-to-end, our RL-based approach directly optimizes the objective function and estimates the instance similarities more precisely. The experimental results show that our method outperforms state-of-the-art competitors with a higher classification accuracy in both synthetic and real examples.

****

Uncovering the Effectiveness of Calibration on Open Intent Classification

Hyejin Won,Kyung Ho Park

Open intent classification aims to simultaneously identify known and unknown intents, and it is one of the challenging tasks in modern dialogue systems. While prior approaches are based on known intent classifiers trained under the cross-entropy loss, we presume this loss function yields a representation overly biased to the known intents; thus, it negatively impacts identifying unknown intents. In this study, we propose a novel open intent classification approach that utilizes model calibration into the previously-proposed state-of-the-art. We empirically examine that simply changing a learning objective in a more calibrated manner outperforms the past state-of-the-art. We further excavate that the underlying reason behind calibrated classifier's supremacy derives from the high-level layers of the deep neural networks. We also discover that our approach is robust to harsh settings where few training samples per class exist. Consequentially, we expect our findings and takeaways to exhibit practical guidelines of open intent classification, thus helping to inform future model design choices.

****

PMixUp: Simultaneous Utilization of Part-of-Speech Replacement and Feature Space Interpolation for Text Data Augmentation

Hyeon Soo Kim,Hyejin Won,Kyung Ho Park

Data augmentation has become a de facto technique in various NLP tasks to overcome the lack of a large-scale, qualified training set. The previous studies presented several data augmentation methods, such as replacing tokens with synonyms or interpolating feature space of given text input. While they are known to be convenient and promising, several limits exist. First, prior studies simply treated topic classification and sentiment analysis under the same category of text classification while we presume they have distinct characteristics. Second, previously-proposed replacement-based methods bear several improvement avenues as they utilize heuristics or statistical approaches for choosing synonyms. Lastly, while the feature space interpolation method achieved current state-of-the-art, prior studies have not comprehensively utilized it with replacement-based methods. To mitigate these drawbacks, we first analyzed which POS tags are important in each text classification task, and resulted that nouns are essential to topic classification, while sentiment analysis regards verbs and adjectives as important POS information. Contrary to the aforementioned analysis, we discover that augmenting verbs and adjective tokens commonly improves text classification performance regardless of its type. Lastly, we propose PMixUp, a novel data augmentation strategy that simultaneously utilizes replacement-based and feature space interpolation methods. We examine that they are new state-of-the-art in nine public benchmark settings, especially under the few training samples.

****

SDT: Specific Domain Training in Domain Generalization

Saeed Karimi,Hamdi Dibeklioglu

Domain generalization (DG) methods aim to achieve generalizability to an unseen target domain by using only training data from the source domains. Although there has been a growing interest to learn from multiple training domains by applying different types of invariance across those domains, the improvements compared to empirical risk minimization (ERM) are almost negligible under controlled evaluation protocols. In this paper, we demonstrate that the disentanglement of spurious and invariant features is a tough task in standard training, since ERM simply minimize the loss and does not exploit invariance among domains. To address the issue, we introduce a simple yet effective method called specific domain training (SDT), which intensifies the trace of spurious features and make them more

discernible and exploit masking strategy to decrease their effect. We provide a theoretical and experimental evidence to show the effectiveness of SDT for out-of-distribution generalization. Notably, SDT outperforms previous state of the art \citet{cha2021swad} in DomainNet benchmarks 0.2pp in average. Furthermore, SDT improves accuracy of some domains such as Sketch in PACS, SUN09 in VLCS and L100 in TerraIncognita by clear margins 2.5pp, 3.4pp, and 5.4pp respectively.
****

Lossy Compression with Gaussian Diffusion
Lucas Theis,Tim Salimans,Matthew Douglas Hoffman,Fabian Mentzer

We consider a novel lossy compression approach based on unconditional diffusion generative models, which we call DiffC. Unlike modern compression schemes which rely on transform coding and quantization to restrict the transmitted information, DiffC relies on the efficient communication of pixels corrupted by Gaussian noise. We implement a proof of concept and find that it works surprisingly well despite the lack of an encoder transform, outperforming the state-of-the-art generative compression method HiFiC on ImageNet 64x64. DiffC only uses a single model to encode and denoise corrupted pixels at arbitrary bitrates. The approach further provides support for progressive coding, that is, decoding from partial bit streams. We perform a rate-distortion analysis to gain a deeper understanding of its performance, providing analytical results for multivariate Gaussian data as well as theoretic bounds for general distributions. Furthermore, we prove that a flow-based reconstruction achieves a 3 dB gain over ancestral sampling at high bitrates.
****

Score-Based Graph Generative Modeling with Self-Guided Latent Diffusion
Ling Yang,Zhilong Zhang,Wentao Zhang,Shenda Hong

Graph generation is a fundamental task in machine learning, and it is critical for numerous real-world applications, biomedical discovery and social science. Existing diffusion-based graph generation methods have two limitations: (i) they conduct diffusion process directly in complex graph space (i.e., node feature, adjacency matrix, or both), resulting in hard optimization with network evaluations; (ii) they usually neglect to sufficiently cover the whole distribution of target unlabeled graph set and thus fail to make semantic controllable generation. In this paper, we first propose a unified latent-based graph generative framework, Score-Based Graph Generative Model (SGGM), powered by Self-Guided Latent Diffusion (SLD) to address both limitations. Specifically, we pretrain a variational graph autoencoder to map raw graph of high-dimensional discrete space to low-dimensional topology-injected latent space, and apply score-based generative model there, yielding a smoother, faster and more expressive graph generation procedure. To sufficiently cover the whole semantical distribution of unlabeled graph set, we propose SLD to make controllable self-guidance of the sample generation with gradients from the designed assigning function towards the hierarchical pseudo label, produced by iteratively clustering on the latent embeddings. In addition, we conduct periodic update on the pseudo label in training process to achieve mutual adaptation between self-guidance and score-based generation. Experiments show that our SGGM powered by SLD outperforms previous graph generation baselines on both generic and molecular graph datasets, demonstrating the generality and extensibility along with further theoretical proofs.
****

Gradient-Informed Quality Diversity for the Illumination of Discrete Spaces
Guillaume Richard,Raphael Boige,Jérémie DONA,Antoine Cully,Thomas PIERROT

Quality Diversity (QD) algorithms have been proposed to search for a large collection of both diverse and high-performing solutions instead of a single set of local optima. While early QD algorithms view the objective and descriptor functions as black-box functions, novel tools have been introduced to use gradient information to accelerate the search and improve overall performance of those algorithms over continuous input spaces. However a broad range of applications involve discrete spaces, such as drug discovery or image generation. Exploring those spaces is challenging as they are combinatorially large and gradients cannot be used in the same manner as in continuous spaces. We introduce MAP-Elites with a Gr adient-Informed Discrete Emitter (ME-GIDE), which extends QD optimisation with differentiable functions over discrete search spaces. ME-GIDE leverages the gradient information of the objective and descriptor functions with respect to its discrete inputs to propose gradient-informed updates that guide the search towards a diverse set of high quality solutions. We evaluate our method on challenging benchmarks including protein design and discrete latent space illumination and find that our method outperforms state-of-the-art QD algorithms in all benchmarks.

****

Deep Generative Wasserstein Gradient Flows
Alvin Heng,Abdul Fatir Ansari,Harold Soh

Deep generative modeling is a rapidly-advancing field with a wealth of modeling choices developed in the past decades. Amongst them, Wasserstein gradient flows (WGF) are a powerful and theoretically rich class of methods.

However, their applications to high-dimensional distributions remain relatively underexplored. In this paper, we present Deep Generative Wasserstein Gradient Flows (DGGF), which constructs a WGF between two distributions by minimizing the entropy-regularized f-divergence. We demonstrate how to train a deep density ratio estimator that is required for the WGF and apply it to the task of generative modeling. Experiments demonstrate that DGGF is able to synthesize high-fidelity images of resolutions up to 128×128, directly in data space. We demonstrate that DGGF has an interpretable diagnostic of sample quality by naturally estimating the KL divergence throughout the gradient flow. Finally, we show DGGF's modularity by composition with external density ratio estimators for conditional generation, as well as for unpaired image-to-image translation with no modifications to the framework.

****

Linear Scalarization for Byzantine-Robust Learning on non-IID data
Latifa Errami,El houcine Bergou

In this work we study the problem of Byzantine-robust learning when data among clients is heterogeneous. We focus on poisoning attacks targeting the convergence of SGD. Although this problem has received great attention; the main Byzantine defenses rely on the IID assumption causing them to fail when data distribution is non-IID even with no attack.

We propose the use of Linear Scalarization (LS) as an enhancing method to enable current defenses to circumvent Byzantine attacks in the non-IID setting. The LS method is based on the incorporation of a trade-off vector that penalizes the suspected malicious clients.

Empirical analysis corroborates that the proposed LS variants are viable in the IID setting. For mild to strong non-IID data splits, LS is either comparable or outperforming current approaches under state-of-the-art Byzantine attack scenarios.

****

Where to Go Next for Recommender Systems? ID- vs. Modality-based recommender models revisited
Zheng Yuan,Fajie Yuan,Yu Song,youhua Li,Fei Yang,Yunzhu Pan

Recommender models that utilize unique identities (IDs for short) to represent distinct users and items have been the state-of-the-arts and dominating the recommender system (RS) literature for over a decade. In parallel, the pre-trained modality encoders, such as BERT and ResNet, are becoming increasingly powerful in modeling raw modality features, e.g., text and images. In light of this, a natural question arises: whether the modality (a.k.a, content) only based recommender models (MoRec) can exceed or be on par with the ID-only based models (IDRec) when item modality features are available? In fact, this question had been answered once a decade ago, when IDRec beat MoRec with strong advantages in terms of both recommendation accuracy and efficiency.

We aim to revisit this `old' question and systematically study MoRec from several aspects. Specifically, we study several sub-questions: (i) which recommender paradigm, MoRec or IDRec, performs best in various practical scenarios, including regular, cold and new item scenarios? does this hold for items with different modality features? (ii) will MoRec benefit from the latest technical advances in corresponding communities, for example, natural language processing and computer vision? (iii) what is an effective way to leverage item modality representations, freezing them or adapting them by fine-tuning on new data? (iv) are there any other factors that affect the efficacy of MoRec. To answer these questions, we conduct rigorous experiments for item recommendations with two popular modalities, i.e., text and vision. We provide empirical evidence that MoRec with standard end-to-end training is highly competitive and even exceeds IDRec in some cases. Many of our observations imply that the dominance of IDRec in terms of recommendation accuracy does not hold well when items' raw modality features are available. We promise to release all related codes & datasets upon acceptance.
****

Optimising 2D Pose Representation: Improving Accuracy, Stability and Generalisability inUnsupervised 2D-3D Human Pose Estimation
Peter Timothy David Hardy,Srinandan Dasmahapatra,Hansung Kim
This paper addresses the problem of 2D pose representation during unsupervised 2D to 3D pose lifting to improve the accuracy, stability and generalisability of 3D human pose estimation (HPE) models. All unsupervised 2D-3D HPE approaches provide the entire 2D kinematic skeleton to a model during training. We argue that this is sub-optimal and disruptive as long-range correlations are induced between independent 2D key points and predicted 3D ordinates during training. To this end, we conduct the following study. With a maximum architecture capacity of 6 residual blocks, we evaluate the performance of 5 models which each represent a 2D pose differently during the adversarial unsupervised 2D-3D HPE process. Additionally, we show the correlations between 2D key points which are learned during the training process, highlighting the unintuitive correlations induced when an entire 2D pose is provided to a lifting model. Our results show that the most optimal representation of a 2D pose is that of two independent segments, the torso and legs, with no shared features between each lifting network. This approach decreased the average error by 20% on the Human3.6M dataset when compared to a model with a near identical parameter count trained on the entire 2D kinematic skeleton. Furthermore, due to the complex nature of adversarial learning, we show how this representation can also improve convergence during training allowing for an optimum result to be obtained more often.
****

Model Obfuscation for Securing Deployed Neural Networks
Mingyi Zhou,Xiang Gao,Jing Wu,John C. Grundy,Xiao Chen,Chunyang Chen,Li Li
More and more edge devices and mobile apps are leveraging deep learning (DL) capabilities. Deploying such models on devices -- referred to as on-device models -- rather than as remote cloud-hosted services, has gained popularity as it avoids transmitting user's data off of the device and for high response time. However, on-device models can be easily attacked, as they can be accessed by unpacking corresponding apps and the model is fully exposed to attackers. Recent studies show that adversaries can easily generate white-box-like attacks for an on-device model or even inverse its training data. To protect on-device models from white-box attacks, we propose a novel technique called model obfuscation. Specifically, model obfuscation hides and obfuscates the key information -- structure, parameters and attributes -- of models by renaming, parameter encapsulation, neural structure obfuscation, shortcut injection, and extra layer injection. We have developed a prototype tool ModelObfuscator to automatically obfuscate on-device TF Lite models. Our experiments show that this proposed approach can dramatically improve model security by significantly increasing the overhead of extracting models' inner information, without increasing the latency of DL models. Our proposed on-device model obfuscation has the potential to be a fundamental technique for on-device model deployment. Our prototype tool is publicly available at https://github.com/AnonymousAuthor000/Code2536.

****

ESP: Exponential Smoothing on Perturbations for Increasing Robustness to Data Corruptions

Weebum Yoo,Hwanjo Yu,Sung Whan Yoon

Despite the great advances in the machine learning field over the past decade, deep learning algorithms are often vulnerable to data corruption in real-world environments. We propose a simple yet efficient data augmentation method named Exponential Smoothing on Perturbations (ESP) that imposes perturbations on training data to enhance a model’s robustness to unforeseen data corruptions. With the perturbation on the input side, the target label of a sample is smoothed with an exponentially decaying confidence level with respect to the size of the perturbation. ESP enforces a contour-like decision boundary that smoothly encompasses the region around inter-class samples. We theoretically show that perturbations in input space can encourage a model to find a flat minimum on the parameter space, which makes a model robust to domain shifts. In the extensive evaluation on common corruption benchmarks including MNIST-C, CIFAR-10/100-C, and Tiny-ImageNet-C, our method improves the robustness of a model both as a standalone method and in conjunction with the previous state-of-the-art augmentation-based methods. ESP is a model-agnostic algorithm in the sense that it is neither model-specific nor data-specific.

****

MultiViz: Towards Visualizing and Understanding Multimodal Models

Paul Pu Liang,Yiwei Lyu,Gunjan Chhablani,Nihal Jain,Zihao Deng,Xingbo Wang,Louis-Philippe Morency,Ruslan Salakhutdinov

The promise of multimodal models for real-world applications has inspired research in visualizing and understanding their internal mechanics with the end goal of empowering stakeholders to visualize model behavior, perform model debugging, and promote trust in machine learning models. However, modern multimodal models are typically black-box neural networks, which makes it challenging to understand their internal mechanics. How can we visualize the internal modeling of multimodal interactions in these models? Our paper aims to fill this gap by proposing MultiViz, a method for analyzing the behavior of multimodal models by scaffolding the problem of interpretability into 4 stages: (1) unimodal importance: how each modality contributes towards downstream modeling and prediction, (2) cross-modal interactions: how different modalities relate with each other, (3) multimodal representations: how unimodal and cross-modal interactions are represented in decision-level features, and (4) multimodal prediction: how decision-level features are composed to make a prediction. MultiViz is designed to operate on diverse modalities, models, tasks, and research areas. Through experiments on 8 trained models across 6 real-world tasks, we show that the complementary stages in MultiViz together enable users to (1) simulate model predictions, (2) assign interpretable concepts to features, (3) perform error analysis on model misclassifications, and (4) use insights from error analysis to debug models. MultiViz is publicly available, will be regularly updated with new interpretation tools and metrics, and welcomes inputs from the community.

****

How Informative is the Approximation Error from Tensor Decomposition for Neural Network Compression?

Jetze Schuurmans,kim batselier,Julian Kooij

Tensor decompositions have been successfully applied to compress neural networks. The compression algorithms using tensor decompositions commonly minimize the approximation error on the weights. Recent work assumes the approximation error on the weights is a proxy for the performance of the model to compress multiple layers and fine-tune the compressed model. Surprisingly, little research has systematically evaluated which approximation errors can be used to make choices regarding the layer, tensor decomposition method, and level of compression. To close this gap, we perform an experimental study to test if this assumption holds across different layers and types of decompositions, and what the effect of fine-tuning is. We include the approximation error on the features resulting from a compressed layer in our analysis to test if this provides a better proxy, as it exp

licitly takes the data into account. We find the approximation error on the weights has a positive correlation with the performance error, before as well as after fine-tuning. Basing the approximation error on the features does not improve the correlation significantly. While scaling the approximation error commonly is used to account for the different sizes of layers, the average correlation across layers is smaller than across all choices (i.e. layers, decompositions, and level of compression) before fine-tuning. When calculating the correlation across the different decompositions, the average rank correlation is larger than across all choices. This means multiple decompositions can be considered for compression and the approximation error can be used to choose between them.
****

DISCO-DANCE: Learning to Discover Skills with Guidance
Hyunseung Kim,Byungkun Lee,Sejik Park,Hojoon Lee,Dongyoon Hwang,Kyushik Min,Jaegul Choo

Unsupervised skill discovery (USD) allows agents to learn diverse and discriminable skills without access to pre-defined rewards,
by maximizing the mutual information (MI) between skills and states reached by each skill.
The most common problem of MI-based skill discovery is insufficient exploration, because each skill is heavily penalized when it deviates from its initial settlement. Recent works introduced an auxiliary reward to encourage the exploration of the agent via maximizing the state's epistemic uncertainty or entropy.
However, we have discovered that the performance of these auxiliary rewards decreases as the environment becomes more challenging. Therefore, we introduce a new unsupervised skill discovery algorithm, skill discovery with guidance (DISCO-DANCE), which (1) selects the guide skill which has the highest potential to reach the unexplored states, (2) guide other skills to follow the guide skill, then (3) the guided skills are diffused to maximize their discriminability in the unexplored states. Empirically, DISCO-DANCE substantially outperforms other USD baselines on challenging environments including two navigation benchmarks and a continuous control benchmark.
****

Architecture-Agnostic Masked Image Modeling -- From ViT back to CNN
Siyuan Li,Di Wu,Fang Wu,Zelin Zang,Lei Shang,Baigui Sun,Xuansong Xie,Stan Z. Li

Masked image modeling (MIM), an emerging self-supervised pre-training method, has shown impressive success across numerous downstream vision tasks with Vision transformers (ViTs). Its underlying idea is simple: a portion of the input image is randomly masked out and then reconstructed via the pre-text task. However, the working principle behind MIM is not well explained, and previous studies insist that MIM primarily works for the Transformer family but is incompatible with CNNs. In this paper, we first study interactions among patches to understand what knowledge is learned and how it is acquired via the MIM task. We observe that MIM essentially teaches the model to learn better middle-order interactions among patches and extract more generalized features. Based on this fact, we propose an Architecture-Agnostic Masked Image Modeling framework (A^{2}MIM), which is compatible with both Transformers and CNNs in a unified way. Extensive experiments on popular benchmarks show that our A^{2}MIM learns better representations without explicit design and endows the backbone model with the stronger capability to transfer to various downstream tasks for both Transformers and CNNs.
****

Blurring Diffusion Models
Emiel Hoogeboom,Tim Salimans

Recently, Rissanen et al., (2022) have presented a new type of diffusion process for generative modeling based on heat dissipation, or blurring, as an alternative to isotropic Gaussian diffusion. Here, we show that blurring can equivalently be defined through a Gaussian diffusion process with non-isotropic noise. In making this connection, we bridge the gap between inverse heat dissipation and denoising diffusion, and we shed light on the inductive bias that results from this modeling choice. Finally, we propose a generalized class of diffusion models that offers the best of both standard Gaussian denoising diffusion and inverse hea t dissipation, which we call Blurring Diffusion Models.

****

BrGANs: Stabilizing GANs' Training Process with Brownian Motion Control

Tianjiao Luo,Ziyu Zhu,Gabriele Oliaro,Jun Zhu,Zhidong Deng

The training process of generative adversarial networks (GANs) is unstable and does not converge globally. In this paper, we propose a universal higher order noise based control called Brownian Motion Control (BMC) that is invariant to GANs frameworks so that the training process of GANs is exponential stable almost surely. Specifically, starting with the prototypical case of Dirac-GANs, we design a BMC and propose Dirac-BrGANs that retrieve exactly the same but reachable optimal equilibrium regardless of GANs' framework. The optimal equilibrium of our Dirac-BrGANs' training system is globally unique and always exists. Furthermore, the training process of Dirac-BrGANs achieve exponentially stability almost surely for any arbitrary initial value. Then we extend our BMC to normal GANs' settings and propose BrGANs. We provide numerical experiments showing that our BrGANs effectively stabilizes GANs's training process and obtains state-of-the art performance compared to other stabilizing methods.

****

Hyperbolic Self-paced Learning for Self-supervised Skeleton-based Action Representations

Luca Franco,Paolo Mandica,Bharti Munjal,Fabio Galasso

Self-paced learning has been beneficial for tasks where some initial knowledge is available, such as weakly supervised learning and domain adaptation, to select and order the training sample sequence, from easy to complex. However its applicability remains unexplored in unsupervised learning, whereby the knowledge of the task matures during training.

We propose a novel HYperbolic Self-Paced model (HYSP) for learning skeletonbased action representations. HYSP adopts self-supervision: it uses data augmentations to generate two views of the same sample, and it learns by matching one (named online) to the other (the target). We propose to use hyperbolic uncertainty to determine the algorithmic learning pace, under the assumption that less uncertain samples should be more strongly driving the training, with a larger weight and pace. Hyperbolic uncertainty is a by-product of the adopted hyperbolic neural networks, it matures during training and it comes with no extra cost, compared to the established Euclidean SSL framework counterparts.

When tested on three established skeleton-based action recognition datasets, HYSP outperforms the state-of-the-art on PKU-MMD I, as well as on 2 out of 3 downstream tasks on NTU-60 and NTU-120. Additionally, HYSP only uses positive pairs and bypasses therefore the complex and computationally-demanding mining procedures required for the negatives in contrastive techniques.

Code is available at https://github.com/paolomandica/HYSP.

****

Unfair geometries: exactly solvable data model with fairness implications

Stefano Sarao Mannelli,Federica Gerace,Negar Rostamzadeh,Luca Saglietti

Machine learning (ML) may be oblivious to human bias but it is not immune to its perpetuation. Marginalisation and iniquitous group representation are often traceable in the very data used for training, and may be reflected or even enhanced by the learning models.

In the present work, we aim at clarifying the role played by data geometry in the emergence of ML bias. We introduce an exactly solvable high-dimensional model of data imbalance, where parametric control over the many bias-inducing factors allows for an extensive exploration of the bias inheritance mechanism.Through the tools of statistical physics, we analytically characterise the typical properties of learning models trained in this synthetic framework and obtain exact predictions for the observables that are commonly employed for fairness assessment.

Despite the simplicity of the data model, we retrace and unpack typical unfairness behaviour observed on real-world datasets.

We also obtain a detailed analytical characterisation of a class of bias mitigation strategies. We first consider a basic loss-reweighing scheme, which allows for an implicit minimisation of different unfairness metrics, and quantify the in compatibilities between some existing fairness criteria. Then, we consider a novel mitigation strategy based on a matched inference approach, consisting in the introduction of coupled learning models. Our theoretical analysis of this approach shows that the coupled strategy can strike superior fairness-accuracy trade-offs.

****

Understanding Adversarial Transferability in Federated Learning

Yijiang Li,ying gao,Dawn Song,Haohan Wang

With the promises Federated Learning (FL) delivers, various topics regarding its robustness and security issues have been widely studied in recent years: such as the possibility to conduct adversarial attacks (or transferable adversarial attacks) in a while-box setting with full knowledge of the model (or the entire data), or the possibility to conduct poisoning/backdoor attacks during the training process as a malicious client. In this paper, we investigate the robustness and security issues from a different, simpler, but practical setting: a group of malicious clients has impacted the model during training by disguising their identities and acting as benign clients, and only revealing their adversary position after the training to conduct transferable adversarial attacks with their data, which is usually a subset of the data that FL system is trained with. Our aim is to offer a full understanding of the challenges the FL system faces in this setting across a spectrum of configurations. We notice that such an attack is possible, but the federated model is more robust compared with its centralized counterpart when the accuracy on clean images is comparable. Through our study, we hypothesized the robustness is from two factors: the decentralized training on distributed data and the averaging operation. Our work has implications for understanding the robustness of federated learning systems and poses a practical question for federated learning applications.

****

Efficient Offline Policy Optimization with a Learned Model

Zichen Liu,Siyi Li,Wee Sun Lee,Shuicheng YAN,Zhongwen Xu

MuZero Unplugged presents a promising approach for offline policy learning from logged data. It conducts Monte-Carlo Tree Search (MCTS) with a learned model and leverages Reanalyze algorithm to learn purely from offline data. For good performance, MCTS requires accurate learned models and a large number of simulations, thus costing huge computing time. This paper investigates a few hypotheses where MuZero Unplugged may not work well under the offline RL settings, including 1) learning with limited data coverage; 2) learning from offline data of stochastic environments; 3) improperly parameterized models given the offline data; 4) with a low compute budget. We propose to use a regularized one-step look-ahead approach to tackle the above issues. Instead of planning with the expensive MCTS, we use the learned model to construct an advantage estimation based on a one-step rollout. Policy improvements are towards the direction that maximizes the estimated advantage with regularization of the dataset. We conduct extensive empirical studies with BSuite environments to verify the hypotheses and then run our algorithm on the RL Unplugged Atari benchmark. Experimental results show that our proposed approach achieves stable performance even with an inaccurate learned model. On the large-scale Atari benchmark, the proposed method outperforms MuZero Unplugged by 43%. Most significantly, it uses only 5.6% wall-clock time (i.e., 1 hour) compared to MuZero Unplugged (i.e., 17.8 hours) to achieve a 150% IQM normalized score with the same hardware and software stacks. Our implementation is open-sourced at https://github.com/sail-sg/rosmo.

****

New Insights for the Stability-Plasticity Dilemma in Online Continual Learning

Dahuin Jung,Dongjin Lee,Sunwon Hong,Hyemi Jang,Ho Bae,Sungroh Yoon

The aim of continual learning is to learn new tasks continuously (i.e., plasticity) without forgetting previously learned knowledge from old tasks (i.e., stability). In the scenario of online continual learning, wherein data comes strictly in a streaming manner, the plasticity of online continual learning is more vulnerable than offline continual learning because the training signal that can be obtained from a single data point is limited. To overcome the stability-plasticity dilemma in online continual learning, we propose an online continual learning framework named multi-scale feature adaptation network (MuFAN) that utilizes a richer context encoding extracted from different levels of a pre-trained network. Additionally, we introduce a novel structure-wise distillation loss and replace the commonly used batch normalization layer with a newly proposed stability-plasticity normalization module to train MuFAN that simultaneously maintains high plasticity and stability. MuFAN outperforms other state-of-the-art continual learning methods on the SVHN, CIFAR100, miniImageNet, and CORe50 datasets. Extensive experiments and ablation studies validate the significance and scalability of each proposed component: 1) multi-scale feature maps from a pre-trained encoder, 2) the structure-wise distillation loss, and 3) the stability-plasticity normalization module in MuFAN. Code is publicly available at https://github.com/whitesnowdrop/MuFAN.

****

MixPro: Data Augmentation with MaskMix and Progressive Attention Labeling for Vision Transformer

Qihao Zhao,Yangyu Huang,Wei Hu,Fan Zhang,Jun Liu

The recently proposed data augmentation TransMix employs attention labels to help visual transformers (ViT) achieve better robustness and performance. However, TransMix is deficient in two aspects: 1) The image cropping method of TransMix may not be suitable for vision transformer. 2) At the early stage of training, the model produces unreliable attention maps. TransMix uses unreliable attention maps to compute mixed attention labels that can affect the model. To address the aforementioned issues, we propose MaskMix and Progressive Attention Labeling (PAL) in image and label space, respectively. In detail, from the perspective of image space, we design MaskMix, which mixes two images based on a patch-like grid mask. In particular, the size of each mask patch is adjustable and is a multiple of the image patch size, which ensures each image patch comes from only one image and contains more global contents. From the perspective of label space, we design PAL, which utilizes a progressive factor to dynamically re-weight the attention weights of the mixed attention label. Finally, we combine MaskMix and Progressive Attention Labeling as our new data augmentation method, named MixPro. The experimental results show that our method can improve various ViT-based models at scales on ImageNet classification (73.8% top-1 accuracy based on DeiT-T for 300 epochs). After being pre-trained with MixPro on ImageNet, the ViT-based models also demonstrate better transferability to semantic segmentation, object detection, and instance segmentation. Furthermore, compared to TransMix, MixPro also shows stronger robustness on several benchmarks.

****

StyleMorph: Disentangled 3D-Aware Image Synthesis with a 3D Morphable StyleGAN

Eric-Tuan Le,Edward Bartrum,Iasonas Kokkinos

We introduce StyleMorph, a 3D-aware generative model that disentangles 3D shape, camera pose, object appearance, and background appearance for high quality image synthesis. We account for shape variability by morphing a canonical 3D object template, effectively learning a 3D morphable model in an entirely unsupervised manner through backprop. We chain 3D morphable modelling with deferred neural rendering by performing an implicit surface rendering of “Template Object Coordinates” (TOCS), which can be understood as an unsupervised counterpart to UV maps. This provides a detailed 2D TOCS map signal that reflects the compounded geometric effects of non-rigid shape variation, camera pose, and perspective projection. We combine 2D TOCS maps with an independent appearance code to condition a StyleGAN-based deferred neural rendering (DNR) network for foreground image (object) synthesis; we use a separate code for background synthesis and do late fusion to deliver the final result. We show competitive synthesis results on 4 datasets (FFHQ faces, AFHQ Cats, Dogs, Wild), while achieving the joint disentanglement of shape, pose, object and background texture.

****

Accelerated Riemannian Optimization: Handling Constraints to Bound Geometric Penalties

David Martínez-Rubio,Sebastian Pokutta

We propose a globally-accelerated, first-order method for the optimization of smooth and (strongly or not) geodesically-convex functions in Hadamard manifolds. Our algorithm enjoys the same convergence rates as Nesterov's accelerated gradient descent, up to a multiplicative geometric penalty and log factors.

Crucially, we can enforce our method to stay within a compact set we define. Prior fully accelerated works resort to assuming that the iterates of their algorithms stay in some pre-specified compact set, except for two previous methods, whose applicability is limited to local optimization and to spaces of constant curvature, respectively. Achieving global and general Riemannian acceleration without iterates assumptively staying in the feasible set was asked as an open question in (Kim & Yang, 2022), which we solve for Hadamard manifolds.

In our solution, we show that we can use a linearly convergent algorithm for constrained strongly g-convex smooth problems to implement a Riemannian inexact proximal point operator that we use as a subroutine, which is of independent interest.

****

Searching Lottery Tickets in Graph Neural Networks: A Dual Perspective

Kun Wang,Yuxuan Liang,Pengkun Wang,Xu Wang,Pengfei Gu,Junfeng Fang,Yang Wang

Graph Neural Networks (GNNs) have shown great promise in various graph learning tasks. However, the computational overheads of fitting GNNs to large-scale graphs grow rapidly, posing obstacles to GNNs from scaling up to real-world applications. To tackle this issue, Graph Lottery Ticket (GLT) hypothesis articulates that there always exists a sparse subnetwork/subgraph with admirable performance in GNNs with random initialization. Such a pair of core subgraph and sparse subnetwork (called graph lottery tickets) can be uncovered by iteratively applying a novel sparsification method. While GLT provides new insights for GNN compression, it requires a full pretraining process to obtain graph lottery tickets, which is not universal and friendly to real-world applications. Moreover, the graph sparsification in GLT utilizes sampling techniques, which may result in massive information loss and aggregation failure. In this paper, we explore the searching of graph lottery tickets from a complementary perspective -- transforming a random ticket into a graph lottery ticket, which allows us to more comprehensively explore the relationships between the original network/graph and their sparse counterpart. To achieve this, we propose regularization-based network pruning and hierarchical graph sparsification, leading to our Dual Graph Lottery Ticket (DGLT) framework for a joint sparsification of network and graph. Compared to GLT, our DGLT helps achieve a triple-win situation of graph lottery tickets with high sparsity, admirable performance, and good explainability. More importantly, we rigorously prove that our model can eliminate noise and maintain reliable information in substructures using the graph information bottleneck theory. Extensive experimental results on various graph-related tasks validate the effectiveness of our framework.

****

Video Scene Graph Generation from Single-Frame Weak Supervision

Siqi Chen,Jun Xiao,Long Chen

Video scene graph generation (VidSGG) aims to generate a sequence of graph-structure representations for the given video. However, all existing VidSGG methods are fully-supervised, i.e., they need dense and costly manual annotations. In this paper, we propose the first weakly-supervised VidSGG task with only single-frame weak supervision: SF-VidSGG. By ``weakly-supervised", we mean that SF-VidSGG relaxes the training supervision from two different levels: 1) It only provides single-frame annotations instead of all-frame annotations. 2) The single-frame ground-truth annotation is still a weak image SGG annotation, i.e., an unlocalized scene graph. To solve this new task, we also propose a novel Pseudo Label Assignment based method, dubbed as PLA. PLA is a two-stage method, which generates pseudo visual relation annotations for the given video at the first stage, and then trains a fully-supervised VidSGG model with these pseudo labels. Specifically, PLA consists of three modules: an object PLA module, a predicate PLA module, and a future predicate prediction (FPP) module. Firstly, in the object PLA, we localize all objects for every frame. Then, in the predicate PLA, we design two di fferent teachers to assign pseudo predicate labels. Lastly, in the FPP module, we fusion these two predicate pseudo labels by the regularity of relation transition in videos. Extensive ablations and results on the benchmark Action Genome have demonstrated the effectiveness of our PLA.

****

Planning With Uncertainty: Deep Exploration in Model-Based Reinforcement Learning

Yaniv Oren,Matthijs T. J. Spaan,Wendelin Boehmer

Deep model-based reinforcement learning has shown super-human performance in many challenging domains. Low sample efficiency and limited exploration remain however as leading obstacles in the field. In this paper, we demonstrate deep exploration in model-based RL by incorporating epistemic uncertainty into planning trees, circumventing the standard approach of propagating uncertainty through value learning. We evaluate this approach with the state of the art model-based RL algorithm MuZero, and extend its training process to stabilize learning from explicitly-exploratory decisions. Our results demonstrate that planning with uncertainty is able to achieve effective deep exploration with standard uncertainty estimation mechanisms, and with it significant gains in sample efficiency.

****

Unsupervised visualization of image datasets using contrastive learning

Niklas Böhm,Philipp Berens,Dmitry Kobak

Visualization methods based on the nearest neighbor graph, such as t-SNE or UMAP, are widely used for visualizing high-dimensional data. Yet, these approaches only produce meaningful results if the nearest neighbors themselves are meaningful. For images represented in pixel space this is not the case, as distances in pixel space are often not capturing our sense of similarity and therefore neighbors are not semantically close. This problem can be circumvented by self-supervised approaches based on contrastive learning, such as SimCLR, relying on data augmentation to generate implicit neighbors, but these methods do not produce two-dimensional embeddings suitable for visualization. Here, we present a new method, called t-SimCNE, for unsupervised visualization of image data. T-SimCNE combines ideas from contrastive learning and neighbor embeddings, and trains a parametric mapping from the high-dimensional pixel space into two dimensions. We show that the resulting 2D embeddings achieve classification accuracy comparable to the state-of-the-art high-dimensional SimCLR representations, thus faithfully capturing semantic relationships. Using t-SimCNE, we obtain informative visualizations of the CIFAR-10 and CIFAR-100 datasets, showing rich cluster structure and highlighting artifacts and outliers.

****

Contrastive Consistent Representation Distillation

Shipeng Fu,haoran Yang,Xiaomin Yang

The combination of knowledge distillation with contrastive learning has great potential to distill structural knowledge. Most of the contrastive-learning-based distillation methods treat the entire training dataset as the memory bank and maintain two memory banks, one for the student and one for the teacher. Besides, the representations in the two memory banks are updated in a momentum manner, leading to representation inconsistency. In this work, we propose Contrastive Consistent Representation Distillation (CoCoRD) to provide consistent representations for efficient contrastive-learning-based distillation. Instead of momentum-updating the cached representations, CoCoRD updates the encoders in a momentum manner. Specifically, the teacher is equipped with a momentum-updated projection head to generate consistent representations. These teacher representations are cached in a fixed-size queue which serves as the only memory bank in CoCoRD and is significantly smaller than the entire training dataset. Additionally, a slow-moving student, implemented as a momentum-based moving average of the student, is built to facilitate contrastive learning. CoCoRD, which utilizes only one memory bank and much fewer negative keys, provides highly competitive results under typical teacher-student settings. On ImageNet, CoCoRD-distilled ResNet50 outperforms the teacher ResNet101 by 0.2% top-1 accuracy. Furthermore, in PASCAL VOC and COCO detection, the detectors whose backbones are initialized by CoCoRD-distilled m odels exhibit considerable performance improvements.

****

PowerQuant: Automorphism Search for Non-Uniform Quantization

Edouard YVINEC,Arnaud Dapogny,Matthieu Cord,Kevin Bailly

Deep neural networks (DNNs) are nowadays ubiquitous in many domains such as computer vision. However, due to their high latency, the deployment of DNNs hinges on the development of compression techniques such as quantization which consists in lowering the number of bits used to encode the weights and activations. Growing concerns for privacy and security have motivated the development of data-free techniques, at the expanse of accuracy. In this paper, we identity the uniformity of the quantization operator as a limitation of existing approaches, and propose a data-free non-uniform method. More specifically, we argue that to be readily usable without dedicated hardware and implementation, non-uniform quantization shall not change the nature of the mathematical operations performed by the DNN. This leads to search among the continuous automorphisms of $(\mathbb{R}_+^*,\times)$, which boils down to the power functions defined by their exponent. To find this parameter, we propose to optimize the reconstruction error of each layer: in particular, we show that this procedure is locally convex and admits a unique solution. At inference time, we show that our approach, dubbed PowerQuant, only require simple modifications in the quantized DNN activation functions. As such, with only negligible overhead, it significantly outperforms existing methods in a variety of configurations.

****

CLEEGN: A Convolutional Neural Network for Plug-and-Play Automatic EEG Reconstruction

Pin-Hua Lai,Wei-Chun Yang,Hsiang-Chieh Tsou,Chun-Shu Wei

Human electroencephalography (EEG) is a brain monitoring modality that senses cortical neuroelectrophysiological activity in high-temporal resolution. One of the greatest challenges posed in applications of EEG is the unstable signal quality susceptible to inevitable artifacts during recordings. To date, most existing techniques for EEG artifact removal and reconstruction are applicable to offline analysis solely, or require individualized training data to facilitate online reconstruction. We have proposed CLEEGN, a novel convolutional neural network for plug-and-play automatic EEG reconstruction. CLEEGN is based on a subject-independent pre-trained model using existing data and can operate on a new user without any further calibration. The performance of CLEEGN was validated using multiple evaluations including waveform observation, reconstruction error assessment, and decoding accuracy on well-studied labeled datasets. The results of simulated online validation suggest that, even without any calibration, CLEEGN can largely preserve inherent brain activity and outperforms leading online/offline artifact removal methods in the decoding accuracy of reconstructed EEG data. In addition, visualization of model parameters and latent features exhibit the model behavior and reveal explainable insights related to existing knowledge of neuroscience. We foresee pervasive applications of CLEEGN in prospective works of online plug-and-play EEG decoding and analysis.

****

On Uni-modal Feature Learning in Multi-modal Learning

Chenzhuang Du,Jiaye Teng,Tingle Li,Yichen Liu,Tianyuan Yuan,Yue Wang,Yang Yuan,Hang Zhao

We abstract the features of multi-modal data into 1) uni-modal features, which can be learned from uni-modal training, and 2) paired features, which can only be learned from cross-modal interaction. Multi-modal joint training is expected to benefit from cross-modal interaction on the basis of ensuring uni-modal feature learning. However, recent late-fusion training approaches still suffer from insufficient learning of uni-modal features on each modality and we prove that this phenomenon does hurt the model's generalization ability.

Given a multi-modal task, we propose to choose targeted late-fusion learning method from Uni-Modal Ensemble (UME) and the proposed Uni-Modal Teacher (UMT), according to the distribution of uni-modal and paired features. We demonstrate that, under a simple guiding strategy, we can achieve comparable results to other com

plex late-fusion or intermediate-fusion methods on multi-modal datasets, including VGG-Sound, Kinetics-400, UCF101, and ModelNet40.
****

Unified neural representation model for physical and conceptual spaces
Tatsuya Haga,Yohei Oseki,Tomoki Fukai

The spatial processing system of the brain uses grid-like neural representations (grid cells) for supporting vector-based navigation. Experiments also suggest that neural representations for concepts (concept cells) exist in the human brain, and conceptual inference relies on navigation in conceptual spaces. We propose a unified model called ``disentangled successor information (DSI)'' that explains neural representations for both physical and conceptual spaces. DSI generates grid-like representations in a 2-dimensional space that highly resemble those observed in the brain. Moreover, the same model creates concept-specific representations from linguistic inputs, corresponding to concept cells. Mathematically, DSI vectors approximate value functions for navigation and word vectors obtained by word embedding methods, thus enabling both spatial navigation and conceptual inference based on vector-based calculation. Our results suggest that a single principle can explain computation of physical and conceptual spaces in the human brain.
****

Symbolic Physics Learner: Discovering governing equations via Monte Carlo tree search
Fangzheng Sun,Yang Liu,Jian-Xun Wang,Hao Sun

Nonlinear dynamics is ubiquitous in nature and commonly seen in various science and engineering disciplines. Distilling analytical expressions that govern nonlinear dynamics from limited data remains vital but challenging. To tackle this fundamental issue, we propose a novel Symbolic Physics Learner (SPL) machine to discover the mathematical structure of nonlinear dynamics. The key concept is to interpret mathematical operations and system state variables by computational rules and symbols, establish symbolic reasoning of mathematical formulas via expression trees, and employ a Monte Carlo tree search (MCTS) agent to explore optimal expression trees based on measurement data. The MCTS agent obtains an optimistic selection policy through the traversal of expression trees, featuring the one that maps to the arithmetic expression of underlying physics. Salient features of the proposed framework include search flexibility and enforcement of parsimony for discovered equations. The efficacy and superiority of the SPL machine are demonstrated by numerical examples, compared with state-of-the-art baselines.
****

The Dynamic of Consensus in Deep Networks and the Identification of Noisy Labels
Daniel Shwartz,Uri Stern,Daphna Weinshall

Deep neural networks have incredible capacity and expressibility, and can seemingly memorize any training set. This introduces a problem when training in the presence of noisy labels, as the noisy examples cannot be distinguished from clean examples by the end of training. Recent research has dealt with this challenge by utilizing the fact that deep networks seem to memorize clean examples much earlier than noisy examples. Here we report a new empirical result: for each example, when looking at the time it has been memorized by each model in an ensemble of networks, the diversity seen in noisy examples is much larger than the clean examples. We use this observation to develop a new method for noisy labels filtration. The method is based on a statistics of the data, which captures the differences in ensemble learning dynamics between clean and noisy data. We test our method on three tasks: (i) noise amount estimation; (ii) noise filtration; (iii) supervised classification. We show that our method improves over existing baselines in all three tasks using a variety of datasets, noise models, and noise levels. Aside from its improved performance, our method has two other advantages. (i) Simplicity, which implies that no additional hyperparameters are introduced. (ii) Our method is modular: it does not work in an end-to-end fashion, and can therefore be used to clean a dataset for any other future usage.
****

Efficient block contrastive learning via parameter-free meta-node approximation

Gayan K Kulatilleke,Marius Portmann,Shekhar S. Chandra

Contrastive learning has recently achieved remarkable success in many domains including graphs. However contrastive loss, especially for graphs, requires a large number of negative samples which is unscalable and computationally prohibitive with a quadratic time complexity. Sub-sampling is not optimal and incorrect negative sampling leads to sampling bias. In this work, we propose a meta-node based approximation technique that can (a) proxy all negative combinations (b) in quadratic cluster size time complexity, (c) at graph level, not node level, and (d) exploit graph sparsity. By replacing node-pairs with additive cluster-pairs, we compute the negatives in cluster-time at graph level. The resulting Proxy approximated meta-node Contrastive (PamC) loss, based on simple optimized GPU operations, captures the full set of negatives, yet is efficient with a linear time complexity. By avoiding sampling, we effectively eliminate sample bias. We meet the criterion for larger number of samples, thus achieving block-contrastiveness, which is proven to outperform pair-wise losses. We use learnt soft cluster assignments for the meta-node constriction, and avoid possible heterophily and noise added during edge creation. Theoretically, we show that real world graphs easily satisfy conditions necessary for our approximation. Empirically, we show promising accuracy gains over state-of-the-art graph clustering on 6 benchmarks. Importantly, we gain substantially in efficiency; up to 3x in training time, 1.8x in inference time and over 5x in GPU memory reduction.

****

Attribute Alignment and Enhancement for Generalized Zero-Shot Learning

Nannan Lu,MingKai Qiu

Generalized zero-shot learning (GZSL) aims to recognize both seen and unseen classes, which challenges the generalization ability of a model. In this paper, we propose a novel approach to fully utilize attributes information, referred to as attribute alignment and enhancement (A3E) network. It contains two modules. First, attribute localization (AL) module utilizes the supervision of class attribute vectors to guide visual localization for attributes through the implicit localization capability within the feature extractor, and the visual features corresponding to the attributes (attribute-visual features) are obtained. Second, enhanced attribute scoring (EAS) module employs the supervision of the attribute word vectors (attribute semantics) to project input attribute visual features to attribute semantic space using Graph Attention Network (GAT). Based on the constructed attribute relation graph (ARG), EAS module generates enhanced representation of attributes. Experiments on standard datasets demonstrate that the enhanced attribute representation greatly improves the classification performance, which helps A3E to achieve state-of-the-art performances in both ZSL and GZSL tasks.

****

BAYES RISK CTC: CONTROLLABLE CTC ALIGNMENT IN SEQUENCE-TO-SEQUENCE TASKS

Jinchuan Tian,Brian Yan,Jianwei Yu,CHAO WENG,Dong Yu,Shinji Watanabe

Sequence-to-Sequence (seq2seq) tasks transcribe the input sequence to a target sequence. The Connectionist Temporal Classification (CTC) criterion is widely used in multiple seq2seq tasks. Besides predicting the target sequence, a side product of CTC is to predict the alignment, which is the most probable input-long sequence that specifies a hard aligning relationship between the input and target units. As there are multiple potential aligning sequences (called paths) that are equally considered in CTC formulation, the choice of which path will be most probable and become the predicted alignment is always uncertain. In addition, it is usually observed that the alignment predicted by vanilla CTC will drift compared with its reference and rarely provides practical functionalities. Thus, the motivation of this work is to make the CTC alignment prediction controllable and thus equip CTC with extra functionalities. The Bayes risk CTC (BRCTC) criterion is then proposed in this work, in which a customizable Bayes risk function is adopted to enforce the desired characteristics of the predicted alignment. With the risk function, the BRCTC is a general framework to adopt some customizable preference over the paths in order to concentrate the posterior into a particular subset of the paths. In applications, we explore one particular preference which yields models with the down-sampling ability and reduced inference costs. By using BRCTC with another preference for early emissions, we obtain an improved performance-latency trade-off for online models. Experimentally, the proposed BRCTC reduces the inference cost of offline models by up to 47% without performance degradation and cuts down the overall latency of online systems to an unseen level.

****

A Convergent Single-Loop Algorithm for Relaxation of Gromov-Wasserstein in Graph Data

Jiajin Li,Jianheng Tang,Lemin Kong,Huikang Liu,Jia Li,Anthony Man-Cho So,Jose Blanchet

In this work, we present the Bregman Alternating Projected Gradient (BAPG) method, a single-loop algorithm that offers an approximate solution to the Gromov-Wasserstein (GW) distance.

We introduce a novel relaxation technique that balances accuracy and computational efficiency, albeit with some compromises in the feasibility of the coupling map. Our analysis is based on the observation that the GW problem satisfies the Luo-Tseng error bound condition, which relates to estimating the distance of a point to the critical point set of the GW problem based on the optimality residual.

This observation allows us to provide an approximation bound for the distance between the fixed-point set of BAPG and the critical point set of GW. Moreover, under a mild technical assumption, we can show that BAPG converges to its fixed point set.

The effectiveness of BAPG has been validated through comprehensive numerical experiments in graph alignment and partition tasks, where it outperforms existing methods in terms of both solution quality and wall-clock time.

****

A Hierarchical Hyper-rectangle Mass Model for Fine-grained Entity Typing

SHING HU,Chunyu Li,Bibo Hao

Fine-grained entity typing is the task of detecting types of entities inside a given language text. Entity typing models typically transform entities into vectors in high-dimensional space, hyperbolic space, or add additional context information. However, such spaces or feature transformations are not compatible with modeling types' inter-dependencies and diverse scenarios. We study the ability of the hierarchical hyper-rectangle mass model(hRMM), which represents mentions and types into hyper-rectangle mass(hRM) and thus captures the relationships of ontology into a geometric mass view. Natural language contexts are fed into the encoder and then projected to hyper-rectangle mass embedding(hRME). We find that hRM perfectly depicts features of mentions and types. With further research in hypervolume indicators and adaptive thresholds, performance achieves additional improvement. Experiments show that our approach achieves better performance on several entity typing benchmarks and attains state-of-the-art results on two benchmark datasets.

****

Semi-supervised learning with a principled likelihood from a generative model of data curation

Stoil Krasimirov Ganev,Laurence Aitchison

We currently do not have an understanding of semi-supervised learning (SSL) objectives such as pseudo-labelling and entropy minimization as log-likelihoods, which precludes the development of e.g. Bayesian SSL. Here, we note that benchmark image datasets such as CIFAR-10 are carefully curated, and we formulate SSL objectives as a log-likelihood in a generative model of data curation. We show that SSL objectives, from entropy minimization and pseudo-labelling, to state-of-the-art techniques similar to FixMatch can be understood as lower-bounds on our principled log-likelihood. We are thus able to introduce a Bayesian extension of SSL, which gives considerable improvements over standard SSL in the setting of 40 labelled points on CIFAR-10, with performance of $92.2\pm 0.3\%$ vs 88.6% in the original FixMatch paper. Finally, our theory suggests that SSL is effective in part due to the statistical patterns induced by data curation. This provides a n explanation of past results which show SSL performs better on clean datasets without any ``out of distribution'' examples. Confirming these results we find that SSL gave much larger performance improvements on curated than on uncurated data, using matched curated and uncurated datasets based on Galaxy Zoo 2.

****

Ti-MAE: Self-Supervised Masked Time Series Autoencoders

Zhe Li,Pengyun Wang,Zhongwen Rao,Lujia Pan,Zenglin Xu

Multivariate Time Series forecasting has been an increasingly popular topic in various applications and scenarios. Recently, contrastive learning and Transformer-based models have achieved good performance in many long-term series forecasting tasks. However, there are still several issues in existing methods. First, the training paradigm of contrastive learning and downstream prediction tasks are inconsistent, leading the accuracy of prediction not good enough. Second, existing Transformer-based models which learn similar patterns in historical time series data to predict future values always induces severe distribution shift problems, and does not fully leverage the sequence information compared to self-supervised methods. To address these issues, we propose a novel framework named Ti-MAE, in which the input time series are assumed to follow an integrate distribution. In detail, Ti-MAE randomly masks out embedded time series data and learns an autoencoder to reconstruct them at the point-level. Ti-MAE adopts mask modeling as the auxiliary task rather than contrastive learning and bridges the connection between existing representation learning and generative Transformer-based methods, reducing the difference between upstream and downstream forecasting tasks while maintaining the utilization of original time series data. Experiments on several public real-world datasets demonstrate that our framework of masked autoencoding could learn strong representations directly from the raw data, yielding better performance in time series forecasting and classification tasks. The code will be made public after this paper is accepted.

****

E3Bind: An End-to-End Equivariant Network for Protein-Ligand Docking

Yangtian Zhang,Huiyu Cai,Chence Shi,Jian Tang

In silico prediction of the ligand binding pose to a given protein target is a crucial but challenging task in drug discovery.

This work focuses on blind flexible self-docking, where we aim to predict the positions, orientations and conformations of docked molecules. Traditional physics-based methods usually suffer from inaccurate scoring functions and high inference cost. Recently, data-driven methods based on deep learning techniques are attracting growing interest thanks to their efficiency during inference and promising performance. These methods usually either adopt a two-stage approach by first predicting the distances between proteins and ligands and then generating the final coordinates based on the predicted distances, or directly predicting the global roto-translation of ligands. In this paper, we take a different route. Inspired by the resounding success of AlphaFold2 for protein structure prediction, we propose E3Bind, an end-to-end equivariant network that iteratively updates the ligand pose. E3Bind models the protein-ligand interaction through careful consideration of the geometric constraints in docking and the local context of the binding site. Experiments on standard benchmark datasets demonstrate the superior performance of our end-to-end trainable model compared to traditional and recently-proposed deep learning methods.

****

Improving Model Consistency of Decentralized Federated Learning via Sharpness Aware Minimization and Multiple Gossip Approaches

Yifan Shi,Li Shen,Kang Wei,Yan Sun,Bo Yuan,Xueqian Wang,Dacheng Tao

To mitigate the privacy leakages and reduce the communication burden of Federated Learning (FL), decentralized FL (DFL) discards the central server and each client only communicates with its neighbors in the decentralized communication network. However, existing DFL algorithms tend to feature high inconsistency among local models, which results in severe distribution shifts across clients and inferior performance compared with centralized FL (CFL), especially on heterogeneous data or with sparse connectivity of communication topology.

To alleviate this challenge, we propose two DFL algorithms named DFedSAM and DFedSAM-MGS to improve the performance.
Specifically, DFedSAM leverages gradient perturbation to generate local flatness models via Sharpness Aware Minimization (SAM), which searches for model parameters with uniformly low loss function values.
In addition, DFedSAM-MGS further boosts DFedSAM by adopting the technique of Multiple Gossip Steps (MGS) for a better model consistency, which accelerates the aggregation of local flatness models and better balances the communication complexity and learning performance.
In the theoretical perspective, we present the improved convergence rates $\small \mathcal{O}\big(\frac{1}{T}+\frac{1}{T^2(1-\lambda)^2}\big)$ and $\small \mathcal{O}\big(\frac{1}{T}+\frac{\lambda^Q+1}{T^2(1-\lambda^Q)^2}\big)$ in the stochastic non-convex setting for DFedSAM and DFedSAM-MGS, respectively, where $1-\lambda$ is the spectral gap of the gossip matrix W and Q is the gossip steps in MGS. Meanwhile, we empirically confirm that our methods can achieve competitive performance compared with CFL baselines and outperform existing DFL baselines.

****

VA-DepthNet: A Variational Approach to Single Image Depth Prediction
Ce Liu,Suryansh Kumar,Shuhang Gu,Radu Timofte,Luc Van Gool
We introduce VA-DepthNet, a simple, effective, and accurate deep neural network approach for the single-image depth prediction (SIDP) problem. The proposed approach advocates using classical first-order variational constraints for this problem. While state-of-the-art deep neural network methods for SIDP learn the scene depth from images in a supervised setting, they often overlook the invaluable invariances and priors in the rigid scene space, such as the regularity of the scene. The paper's main contribution is to reveal the benefit of classical and well-founded variational constraints in the neural network design for the SIDP task. It is shown that imposing first-order variational constraints in the scene space together with popular encoder-decoder-based network architecture design provides excellent results for the supervised SIDP task. The imposed first-order variational constraint makes the network aware of the depth gradient in the scene space, i.e., regularity. The paper demonstrates the usefulness of the proposed approach via extensive evaluation and ablation analysis over several benchmark datasets, such as KITTI, NYU Depth V2, and SUN RGB-D. The VA-DepthNet at test time shows considerable improvements in depth prediction accuracy compared to the prior art and is accurate also at high-frequency regions in the scene space. At the time of writing this paper, our method---labeled as VA-DepthNet, when tested on the KITTI depth-prediction evaluation set benchmarks, shows state-of-the-art results, and is the top-performing published approach.
****
Prompt-to-Prompt Image Editing with Cross-Attention Control
Amir Hertz,Ron Mokady,Jay Tenenbaum,Kfir Aberman,Yael Pritch,Daniel Cohen-or
Recent large-scale text-driven synthesis diffusion models have attracted much attention thanks to their remarkable capabilities of generating highly diverse images that follow given text prompts. Therefore, it is only natural to build upon these synthesis models to provide text-driven image editing capabilities. However, Editing is challenging for these generative models, since an innate property of an editing technique is to preserve some content from the original image, while in the text-based models, even a small modification of the text prompt often leads to a completely different outcome. State-of-the-art methods mitigate this by requiring the users to provide a spatial mask to localize the edit, hence, ignoring the original structure and content within the masked region. In this paper, we pursue an intuitive prompt-to-prompt editing framework, where the edits are controlled by text only. We analyze a text-conditioned model in depth and observe that the cross-attention layers are the key to controlling the relation between the spatial layout of the image to each word in the prompt. With this observation, we propose to control the attention maps along the diffusion process. Our approach enables us to monitor the synthesis process by editing the textual prompt only, paving the way to a myriad of caption-based editing applications such

as localized editing by replacing a word, global editing by adding a specification, and even controlling the extent to which a word is reflected in the image. We present our results over diverse images and prompts with different text-to-image models, demonstrating high-quality synthesis and fidelity to the edited prompts.

****

Re-weighting Based Group Fairness Regularization via Classwise Robust Optimization

Sangwon Jung,Taeeon Park,Sanghyuk Chun,Taesup Moon

Many existing group fairness-aware training methods aim to achieve the group fairness by either re-weighting underrepresented groups based on certain rules or using weakly approximated surrogates for the fairness metrics in the objective as regularization terms. Although each of the learning schemes has its own strength in terms of applicability or performance, respectively, it is difficult for any method in the either category to be considered as a gold standard since their successful performances are typically limited to specific cases. To that end, we propose a principled method, dubbed as FairDRO, which unifies the two learning schemes by incorporating a well-justified group fairness metric into the training objective using a classwise distributionally robust optimization (DRO) framework. We then develop an iterative optimization algorithm that minimizes the resulting objective by automatically producing the correct re-weights for each group. Our experiments show that FairDRO is scalable and easily adaptable to diverse applications, and consistently achieves the state-of-the-art performance on several benchmark datasets in terms of the accuracy-fairness trade-off, compared to recent strong baselines.

****

DiffEdit: Diffusion-based semantic image editing with mask guidance

Guillaume Couairon,Jakob Verbeek,Holger Schwenk,Matthieu Cord

Image generation has recently seen tremendous advances, with diffusion models allowing to synthesize convincing images for a large variety of text prompts. In this article, we propose DiffEdit, a method to take advantage of text-conditioned diffusion models for the task of semantic image editing, where the goal is to edit an image based on a text query. Semantic image editing is an extension of image generation, with the additional constraint that the generated image should be as similar as possible to a given input image.

Current editing methods based on diffusion models usually require to provide a mask, making the task much easier by treating it as a conditional inpainting task. In contrast, our main contribution is able to automatically generate a mask highlighting regions of the input image that need to be edited, by contrasting predictions of a diffusion model conditioned on different text prompts. Moreover, we rely on latent inference to preserve content in those regions of interest and show excellent synergies with mask-based diffusion.

DiffEdit achieves state-of-the-art editing performance on ImageNet. In addition, we evaluate semantic image editing in more challenging settings, using images from the COCO dataset as well as text-based generated images.

****

Are More Layers Beneficial to Graph Transformers?

Haiteng Zhao,Shuming Ma,Dongdong Zhang,Zhi-Hong Deng,Furu Wei

Despite that going deep has proven successful in many neural architectures, the existing graph transformers are relatively shallow. In this work, we explore whether more layers are beneficial to graph transformers, and find that current graph transformers suffer from the bottleneck of improving performance by increasing depth. Our further analysis reveals the reason is that deep graph transformers are limited by the vanishing capacity of global attention, restricting the graph transformer from focusing on the critical substructure and obtaining expressive features. To this end, we propose a novel graph transformer model named DeepGraph that explicitly employs substructure tokens in the encoded representation, and applies local attention on related nodes to obtain substructure based attention encoding. Our model enhances the ability of the global attention to focus on substructures and promotes the expressiveness of the representations, addressing

the limitation of self-attention as the graph transformer deepens. Experiments show that our method unblocks the depth limitation of graph transformers and results in state-of-the-art performance across various graph benchmarks with deeper models.
****

Learning Combinatorial Node Labeling Algorithms
Lukas Gianinazzi,Maximilian Fries,Nikoli Dryden,Tal Ben-Nun,Maciej Besta,Torsten Hoefler
We present the combinatorial node labeling framework, which generalizes many prior approaches to solving hard graph optimization problems by supporting problems where solutions consist of arbitrarily many node labels, such as graph coloring. We then introduce a neural network architecture to implement this framework. Our architecture builds on a graph attention network with several inductive biases to improve solution quality and is trained using policy gradient reinforcement learning. We demonstrate our approach on both graph coloring and minimum vertex cover. Our learned heuristics match or outperform classical hand-crafted greedy heuristics and machine learning approaches while taking only seconds on large graphs. We conduct a detailed analysis of the learned heuristics and architecture choices and show that they successfully adapt to different graph structures.
****

Simplicial Hopfield networks
Thomas F Burns,Tomoki Fukai
Hopfield networks are artificial neural networks which store memory patterns on the states of their neurons by choosing recurrent connection weights and update rules such that the energy landscape of the network forms attractors around the memories. How many stable, sufficiently-attracting memory patterns can we store in such a network using N neurons? The answer depends on the choice of weights and update rule. Inspired by setwise connectivity in biology, we extend Hopfield networks by adding setwise connections and embedding these connections in a simplicial complex. Simplicial complexes are higher dimensional analogues of graphs which naturally represent collections of pairwise and setwise relationships. We show that our simplicial Hopfield networks increase memory storage capacity. Surprisingly, even when connections are limited to a small random subset of equivalent size to an all-pairwise network, our networks still outperform their pairwise counterparts. Such scenarios include non-trivial simplicial topology. We also test analogous modern continuous Hopfield networks, offering a potentially promising avenue for improving the attention mechanism in Transformer models.
****

Versatile Neural Processes for Learning Implicit Neural Representations
Zongyu Guo,Cuiling Lan,Zhizheng Zhang,Yan Lu,Zhibo Chen
Representing a signal as a continuous function parameterized by neural network (a.k.a. Implicit Neural Representations, INRs) has attracted increasing attention in recent years. Neural Processes (NPs), which model the distributions over functions conditioned on partial observations (context set), provide a practical solution for fast inference of continuous functions. However, existing NP architectures suffer from inferior modeling capability for complex signals. In this paper, we propose an efficient NP framework dubbed Versatile Neural Processes (VNP), which largely increases the capability of approximating functions. Specifically, we introduce a bottleneck encoder that produces fewer and informative context tokens, relieving the high computational cost while providing high modeling capability. At the decoder side, we hierarchically learn multiple global latent variables that jointly model the global structure and the uncertainty of a function, enabling our model to capture the distribution of complex signals. We demonstrate the effectiveness of the proposed VNP on a variety of tasks involving 1D, 2D and 3D signals. Particularly, our method shows promise in learning accurate INRs w.r.t. a 3D scene without further finetuning.
****

DEEP ACCURATE SOLVER FOR THE GEODESIC PROBLEM
Saar Huberman,Ron Kimmel
A high order accurate deep learning method for computing geodesic distances on s

urfaces is introduced. The proposed method exploits a dynamic programming principle which lends itself to a scheme with quasi-linear computational complexity. We consider two main components for computing distances on surfaces; A numerical solver that locally approximates the distance function and an efficient causal ordering scheme by which points are updated. The quality of the distance approximation is determined by the local solver and is the main focus of this paper. A common approach to compute distances on continuous surfaces is by considering a discretized polygonal mesh approximating the surface, and estimating distances on the polygon. With such an approximation, the exact geodesic distances restricted to the polygon are at most second order accurate with respect to the distances on the corresponding continuous surface. Here, by order of accuracy we refer to the rate of convergence as a function of the average distance between sampled points. To improve the rate of convergence, we consider a neural network based local solver which implicitly approximates the structure of the continuous surface. The proposed solver circumvents the polyhedral representation, by directly consuming sampled mesh vertices for approximation of distances on the sampled continuous surfaces. We provide numerical evidence that the proposed update scheme, with appropriate local numerical support, provides better accuracy compared to the best possible polyhedral approximations and previous learning based methods. We introduce a trained solver which is third order accurate, with quasi-linear complexity in the number of sampled points.
****

PBFormer: Capturing Complex Scene Text Shape with Polynomial Band Transformer
Ruijin Liu,Ning Lu,Dapeng Chen,Cheng LI,Zejian Yuan,Wei Peng
We present PBFormer, an efficient yet powerful scene text detector that unifies the transformer with a novel text shape representation Polynomial Band (PB). The representation has four polynomial curves to fit a text's top, bottom, left, and right sides, which can capture a text with a complex shape by varying polynomial coefficients. PB has appealing features compared with conventional representations: 1) It can model different curvatures with a fixed number of parameters, while polygon-points-based methods need to utilize a different number of points. 2) It can distinguish adjacent or overlapping texts as they have apparent different curve coefficients, while segmentation-based methods suffer from adhesive spatial positions. PBFormer combines the PB with the transformer, which can directly generate smooth text contours sampled from predicted curves without interpolation. To leverage the advantage of PB, PBFormer has a parameter-free cross-scale pixel attention module. The module can enlarge text features and suppress irrelevant areas to benefit from detecting texts with diverse scale variations. Furthermore, PBFormer is trained with a shape-contained loss, which not only enforces the piecewise alignment between the ground truth and the predicted curves but also makes curves' position and shapes consistent with each other. Without bells and whistles about text pre-training, our method is superior to the previous state-of-the-art text detectors on the arbitrary-shaped CTW1500 and Total-Text datasets. Codes will be public.
****

Classically Approximating Variational Quantum Machine Learning with Random Fourier Features
Jonas Landman,Slimane Thabet,Constantin Dalyac,Hela Mhiri,Elham Kashefi
Many applications of quantum computing in the near term rely on variational quantum circuits (VQCs). They have been showcased as a promising model for reaching a quantum advantage in machine learning with current noisy intermediate scale quantum computers (NISQ). It is often believed that the power of VQCs relies on their exponentially large feature space, and extensive works have explored the expressiveness and trainability of VQCs in that regard. In our work, we propose a classical sampling method that can closely approximate most VQCs with Hamiltonian encoding, given only the description of their architecture. It uses the seminal proposal of Random Fourier Features (RFF) and the fact that VQCs can be seen as large Fourier series. We show theoretically and experimentally that models built from exponentially large quantum feature space can be classically reproduced by sampling a few frequencies to build an equivalent low dimensional kernel. Prec

isely, we show that the number of required samples grows favourably with the size of the quantum spectrum. This tool therefore questions the hope for quantum advantage from VQCs in many cases, but conversely helps to narrow the conditions for their potential success. We expect VQCs with various and complex encoding Hamiltonians, or with large input dimension, to become more robust to classical approximations.

****

Distributional Meta-Gradient Reinforcement Learning

Haiyan Yin,Shuicheng YAN,Zhongwen Xu

Meta-gradient reinforcement learning (RL) algorithms have substantially boosted the performance of RL agents by learning an adaptive return. All the existing algorithms adhere to the same reward learning principle, where the adaptive return is simply formulated in the form of expected cumulative rewards, upon which the policy and critic update rules are specified under well-adopted distance metrics. In this paper, we present a novel algorithm that builds on the success of meta-gradient RL algorithms and effectively improves such algorithms by following a simple recipe, i.e., going beyond the expected return to formulate and learn the return in a more expressive form, value distributions. To this end, we first formulate a distributional return that could effectively capture bootstrapping and discounting behaviors over distributions, to form an informative distributional return target in value update. Then we derive an efficient meta update rule to learn the adaptive distributional return with meta-gradients. For empirical evaluation, we first present an illustrative example on a toy two-color grid-world domain, which validates the benefit of learning distributional return over expectation; then we conduct extensive comparisons on a large-scale RL benchmark Atari 2600, where we confirm that our proposed method with distributional return works seamlessly well with the actor-critic framework and leads to state-of-the-art median human normalized score among meta-gradient RL literature.

****

Towards A Unified Policy Abstraction Theory and Representation Learning Approach in Markov Decision Processes

Min Zhang,Hongyao Tang,Jianye HAO,YAN ZHENG

Lying on the heart of intelligent decision-making systems, how policy is represented and optimized is a fundamental problem. The root challenge in this problem is the large scale and the high complexity of policy space, which exacerbates the difficulty of policy learning especially in real-world scenarios. Towards a desirable surrogate policy space, recently policy representation in a low-dimensional latent space has shown its potential in improving both the evaluation and optimization of policy. The key question involved in these studies is by what criterion we should abstract the policy space for desired compression and generalization. However, both the theory on policy abstraction and the methodology on policy representation learning are less studied in the literature. In this work, we make very first efforts to fill up the vacancy. First, we propose a unified policy abstraction theory, containing three types of policy abstraction associated to policy features at different levels. Then, we generalize them to three policy metrics that quantify the distance (i.e., similarity) of policies, for more convenient use in learning policy representation. Further, we propose a policy representation learning approach based on deep metric learning. For the empirical study, we investigate the efficacy of the proposed policy metrics and representations, in characterizing policy difference and conveying policy generalization respectively. Our experiments are conducted in both policy optimization and evaluation problems, containing trust-region policy optimization (TRPO), diversity-guided evolution strategy (DGES) and off-policy evaluation (OPE). Somewhat naturally, the experimental results indicate that there is no a universally optimal abstraction for all downstream learning problems; while the influence-irrelevance policy abstraction can be a generally preferred choice.

****

CENTROID-BASED JOINT REPRESENTATION FOR HUMAN POSE ESTIMATION AND INSTANCE SEGMENTATION

Niaz Ahmad,Jawad Khan,Jeremy Yuhyun Kim,Youngmoon Lee

Joint pose estimation and instance segmentation combines keypoint heatmaps with segmentation masks for multi-person pose and instance-level segmentation. Unlike easy cases with explicit heatmap activation, hard cases with implicit heatmap due to multi-person entanglement, overlap, and occlusions requires joint representation with a segmentation mask in end-to-end training. This paper presents a new centroid-based joint representation method called CENTERFOCUS. It follows a bottom-up paradigm to generate Strong Keypoint Feature Maps for both soft and hard keypoints and improve keypoints detection accuracy as well as the confidence score by introducing KeyCentroids and a Body Heat Map. CENTERFOCUS then uses the high-resolution representation of keypoint as a center of attraction for the pixels in the embedding space to generate MaskCentroid to cluster the pixels to a particular human instance to whom it belongs, even if 70% of the body is occluded. Finally, we propose a new PoseSeg algorithm that collects the feature representation of a 2D human pose and segmentation for the joint structure of the pose and instance segmentation. We then experimentally demonstrate the effectiveness and generalization ability of our system on challenging scenarios such as occlusions, entangled limbs, and overlapping people. The experimental results show the effectiveness of CENTERFOCUS outperforms representative models on the challenging MS COCO and OCHuman benchmarks in terms of both accuracy and runtime performance, Ablation experiments analyze the impact of each component of the system. The code will be released publicly.

****

Pairwise Confidence Difference on Unlabeled Data is Sufficient for Binary Classification

Wei Wang,Lei Feng,Gang Niu,Min-Ling Zhang,Masashi Sugiyama

Learning with confidence labels is an emerging weakly supervised learning paradigm, where training data are equipped with confidence labels instead of exact labels. Positive-confidence (Pconf) classification is a typical learning problem in this context, where we are given only positive data equipped with confidence. However, pointwise confidence may not be accessible in real-world scenarios. In this paper, we dive into a novel weakly supervised learning problem called confidence-difference (ConfDiff) classification. Instead of pointwise confidence, we are given only unlabeled data pairs equipped with confidence difference specifying the difference in the probabilities of being positive. An unbiased risk estimator is derived to tackle the problem, and we show that the estimation error bound achieves the optimal convergence rate. Extensive experiments on benchmark data sets validate the effectiveness of our proposed approaches in leveraging the supervision information of the confidence difference.

****

Emergent Communication with Attention

Ryokan Ri,Ryo Ueda,Jason Naradowsky

To develop computational agents that can better communicate with others using their own emergent language, we endow the agents with an ability to focus their attention on particular concepts in the environment. Humans often understand a thing or scene as a composite of concepts and those concepts are further mapped onto words. We implement this intuition as attention mechanisms in Speaker and Listener agents in a referential game and show attention leads to more compositional and interpretable emergent language. We also demonstrate how attention helps us understand the learned communication protocol by investigating the attention weights associated with each message symbol and the alignment of attention weights between Speaker and Listener agents. Overall, our results suggest that attention is a promising mechanism for developing more human-like emergent language.

****

Discovering Bugs in Vision Models using Off-the-shelf Image Generation and Captioning

Olivia Wiles,Isabela Albuquerque,Sven Gowal

Automatically discovering failures in vision models under real-world settings remains an open challenge. This work shows how off-the-shelf, large-scale, image-to-text and text-to-image models, trained on vast amounts of data, can be leveraged to automatically find such failures. In essence, a conditional text-to-image generative model is used to generate large amounts of synthetic, yet realistic, inputs given a ground-truth label. A captioning model is used to describe misclassified inputs. Descriptions are used in turn to generate more inputs, thereby assessing whether specific descriptions induce more failures than expected. As failures are grounded to natural language, we automatically obtain a high-level, human-interpretable explanation of each failure. We use this pipeline to demonstrate that we can effectively interrogate classifiers trained on ImageNet to find specific failure cases and discover spurious correlations. We also show that we can scale the approach to generate adversarial datasets targeting specific classifier architectures. This work demonstrates the utility of large-scale generative models to automatically discover bugs in vision models in an open-ended manner. We also describe a number of limitations and pitfalls related to this approach.

****

MetaFS: An Effective Wrapper Feature Selection via Meta Learning

Zheyi Pan,Chunyang Li,Songyu Ke,Haoran Xu,Junbo Zhang,Ye Yuan,Yu Zheng

Feature selection is of great importance and applies in lots of fields, such as medical and commercial. Wrapper methods, directly comparing the performance of different feature combinations, are widely used in real-world applications. However, selecting effective features meets the following two main challenges: 1) feature combinations are distributed in a huge discrete space; and 2) efficient and precise combinations evaluation is hard. To tackle these challenges, we propose a novel deep meta-learning-based feature selection framework, termed MetaFS, containing a Feature Subset Sampler (FSS) and a Meta Feature Estimator (MetaFE), which transforms the discrete search space into continuous and adopts meta-learning technique for effective feature selection. Specifically, FSS parameterizes the distribution of discrete search space and applies gradient-based methods to optimize. MetaFE learns the representations of different feature combinations, and dynamically generates unique models without retraining for efficient and precise combination evaluation. We adopt a bi-level optimization strategy to optimize the MetaFS. After optimization, we evaluate multiple feature combinations sampled from the converged distribution (i.e., the condensed search space) and select the optimal one. Finally, we conduct extensive experiments on two datasets, illustrating the superiority of MetaFS over 7 state-of-the-art methods.

****

Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models

Hong Liu,Sang Michael Xie,Zhiyuan Li,Tengyu Ma

Language modeling on large-scale datasets leads to impressive performance gains on various downstream language tasks. The (validation) pre-training loss (or perplexity in autoregressive language modeling) is often used as the evaluation metric when developing language models since the pre-training loss tends to be well-correlated with downstream performance (which is itself difficult to evaluate comprehensively). Contrary to this conventional wisdom, this paper shows that 1) pre-training loss cannot fully explain downstream performance and 2) flatness of the model is well-correlated with downstream performance where pre-training loss is not. On simplified datasets, we identify three ways to produce models with the same (statistically optimal) pre-training loss but different downstream performance: continue pre-training after convergence, increasing the model size, and changing the training algorithm. These experiments demonstrate the existence of implicit bias of pre-training algorithms/optimizers---among models with the same minimal pre-training loss, they implicitly prefer more transferable ones. Toward understanding this implicit bias, we prove that SGD with standard mini-batch noise implicitly prefers flatter minima in language models, and empirically observe a strong correlation between flatness and downstream performance among models with the same minimal pre-training loss. We also prove in a synthetic langua

ge setting that among the models with the minimal pre-training loss, the flattest model transfers to downstream tasks.
****
Signal to Sequence Attention-Based Multiple Instance Network for Segmentation Free Inference of RNA Modifications
Christopher Hendra,Alexandre H. Thiery,Jonathan Goeke
Direct RNA sequencing technology works by allowing long RNA molecules to pass through tiny pores, generating electrical current, called squiggle, that are interpreted as a series of RNA nucleotides through the use of Deep Learning algorithms. The platform has also facilitated computational detection of RNA modifications via machine learning and statistical approaches as they cause detectable shift in the current generated as the modified nucleotides pass through the pores. Nevertheless, since modifications only occur in a handful of positions along the molecules, existing techniques require segmentation of the long squiggle in order to filter off irrelevant signals and this step produces large computational and storage overhead. Inspired by the recent work in vector similarity search, we introduce a segmentation-free approach by utilizing scaled-dot product attention to perform implicit segmentation and feature extraction of raw signals that correspond to sites of interest. We further demonstrate the feasibility of our approach by achieving significant speedup while maintaining competitive performance in m6A detection against existing state-of-the-art methods.
****
Interval-based Offline Policy Evaluation without Sufficient Exploration or Realizability
Kohei Miyaguchi,Hiroshi Kajino
We study the problem of offline policy evaluation (OPE),
where the goal is to estimate the value of given decision-making policy without interacting with the actual environment.
In particular, we consider the interval-based OPE, where the output is an interval rather than a point, indicating the uncertainty of the evaluation.
The interval-based estimation is especially important in OPE since,
when the data coverage is insufficient relative to the complexity of the environmental model,
any OPE method can be biased even with infinite sample size.
In this paper, we characterize the worst case of such irreducible bias, called the *minimax bias*, in terms of the discrepancy between the target policy and the data-sampling distribution,
and show that the marginal-importance-sampling (MIS) estimator achieves the minimax bias with an appropriate importance-weight function.
Motivated with this result, we then propose a new interval-based MIS estimator that asymptotically achieves the minimax bias.
****
A Differential Geometric View and Explainability of GNN on Evolving Graphs
Yazheng Liu,Xi Zhang,Sihong Xie
Graphs are ubiquitous in social networks and biochemistry, where Graph Neural Networks (GNN) are the state-of-the-art models for prediction. Graphs can be evolving and it is vital to formally model and understand how a trained GNN responds to graph evolution. We propose a smooth parameterization of the GNN predicted distributions using axiomatic attribution, where the distributions are on a low dimensional manifold within a high-dimensional embedding space. We exploit the differential geometric viewpoint to model distributional evolution as smooth curves on the manifold. We reparameterize families of curves on the manifold and design a convex optimization problem to find a unique curve that concisely approximates the distributional evolution for human interpretation. Extensive experiments on node classification, link prediction, and graph classification tasks with evolving graphs demonstrate the better sparsity, faithfulness, and intuitiveness of the proposed method over the state-of-the-art methods.
****
$\rm A^2Q$: Aggregation-Aware Quantization for Graph Neural Networks
Zeyu Zhu,Fanrong Li,Zitao Mo,Qinghao Hu,Gang Li,Zejian Liu,Xiaoyao Liang,Jian Ch

eng

As graph data size increases, the vast latency and memory consumption during inference pose a significant challenge to the real-world deployment of Graph Neural Networks (GNNs). While quantization is a powerful approach to reducing GNNs complexity, most previous works on GNNs quantization fail to exploit the unique characteristics of GNNs, suffering from severe accuracy degradation. Through an in-depth analysis of the topology of GNNs, we observe that the topology of the graph leads to significant differences between nodes, and most of the nodes in a graph appear to have a small aggregation value. Motivated by this, in this paper, we propose the Aggregation-Aware mixed-precision Quantization ($\rm A^2Q$) for GNNs, where an appropriate bitwidth is automatically learned and assigned to each node in the graph. To mitigate the vanishing gradient problem caused by sparse connections between nodes, we propose a Local Gradient method to serve the quantization error of the node features as the supervision during training. We also develop a Nearest Neighbor Strategy to deal with the generalization on unseen graphs. Extensive experiments on eight public node-level and graph-level datasets demonstrate the generality and robustness of our proposed method. Compared to the FP32 models, our method can achieve up to $18.8\times$ (i.e., 1.70bits) compression ratio with negligible accuracy degradation. Moreover, compared to the state-of-the-art quantization method, our method can achieve up to 11.4% and 9.5% accuracy improvements on the node-level and graph-level tasks, respectively, and up to $2\times$ speedup on a dedicated hardware accelerator.

****

Multi-Prompt Alignment for Multi-source Unsupervised Domain Adaptation

Haoran Chen,Zuxuan Wu,Yu-Gang Jiang

Most existing methods for multi-source unsupervised domain adaptation (UDA) rely on a common feature encoder to extract domain-invariant features. However, learning such an encoder involves updating the parameters of the entire network, which makes the optimization computationally expensive, particularly when coupled with min-max objectives. Inspired by recent advances in prompt learning that adapts high-capacity deep models for downstream tasks in a computationally economic way, we introduce Multi-Prompt Alignment (MPA), a simple yet efficient two-stage framework for multi-source UDA. Given a source and target domain pair, MPA first trains an individual prompt to minimize the domain gap through a contrastive loss, while tuning only a small set of parameters. Then, MPA derives a low-dimensional latent space through an auto-encoding process that maximizes the agreement of multiple learned prompts. The resulting embedding further facilitates generalization to unseen domains. Extensive experiments show that our method achieves state-of-the-art results on popular benchmark datasets while requiring substantially fewer tunable parameters. To the best of our knowledge, we are the first to apply prompt learning to the multi-source UDA problem and our method achieves the highest reported average accuracy of 54.1% on DomainNet, the most challenging UDA dataset to date, with only 15.9M parameters trained. More importantly, we demonstrate that the learned embedding space can be easily adapted to novel unseen domains.

****

Clean-image Backdoor: Attacking Multi-label Models with Poisoned Labels Only

Kangjie Chen,Xiaoxuan Lou,Guowen Xu,Jiwei Li,Tianwei Zhang

Multi-label models have been widely used in various applications including image annotation and object detection. The fly in the ointment is its inherent vulnerability to backdoor attacks due to the adoption of deep learning techniques. However, all existing backdoor attacks exclusively require to modify training inputs (e.g., images), which may be impractical in real-world applications. In this paper, we aim to break this wall and propose the first clean-image backdoor attack, which only poisons the training labels without touching the training samples. Our key insight is that in a multi-label learning task, the adversary can just manipulate the annotations of training samples consisting of a specific set of classes to activate the backdoor. We design a novel trigger exploration method to find convert and effective triggers to enhance the attack performance. We also propose three target label selection strategies to achieve different goals. Expe rimental results indicate that our clean-image backdoor can achieve a 98% attack success rate while preserving the model's functionality on the benign inputs. Besides, the proposed clean-image backdoor can evade existing state-of-the-art defenses.

****

Dense Correlation Fields for Motion Modeling in Action Recognition

Yichen Qian,Junyan Wang,Xiuyu Sun

The challenge of action recognition is to capture reasoning motion information. Compared to spatial convolution for appearance, the temporal component provides an additional (and important) clue for motion modeling, as a number of actions can be reliably recognized based on the motion information. In this paper, we present an effective and interpretable module, Dense Correlation Fields (DCF), which builds up dense visual correlation volumes at the feature level to model different motion patterns explicitly. To achieve this goal, we rely on a spatially hierarchical architecture that preserves both fine local information provided in the lower layer and the high-level semantic information from the deeper layer. Our method fuses spatial hierarchical correlation and temporal long-term correlation, which is better suited for small objects and large displacements. This module is extensible and can be plugged into many backbone architectures to accurately predict object interactions in the video. DCF shows consistent improvements over 2D CNNs and 3D CNNs baseline networks with 3.7% and 3.0% gains respectively on the standard video action benchmark of SSV1.

****

Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top

Eduard Gorbunov,Samuel Horváth,Peter Richtárik,Gauthier Gidel

Byzantine-robustness has been gaining a lot of attention due to the growth of the interest in collaborative and federated learning. However, many fruitful directions, such as the usage of variance reduction for achieving robustness and communication compression for reducing communication costs, remain weakly explored in the field. This work addresses this gap and proposes Byz-VR-MARINA -- a new Byzantine-tolerant method with variance reduction and compression. A key message of our paper is that variance reduction is key to fighting Byzantine workers more effectively. At the same time, communication compression is a bonus that makes the process more communication efficient. We derive theoretical convergence guarantees for Byz-VR-MARINA outperforming previous state-of-the-art for general non-convex and Polyak-Lojasiewicz loss functions. Unlike the concurrent Byzantine-robust methods with variance reduction and/or compression, our complexity results are tight and do not rely on restrictive assumptions such as boundedness of the gradients or limited compression. Moreover, we provide the first analysis of a Byzantine-tolerant method supporting non-uniform sampling of stochastic gradients. Numerical experiments corroborate our theoretical findings.

****

What's Behind the Mask: Estimating Uncertainty in Image-to-Image Problems

Gilad Kutiel,Regev Cohen,Michael Elad,Daniel Freedman

Estimating uncertainty in image-to-image networks is an important task, particularly as such networks are being increasingly deployed in the biological and medical imaging realms. In this paper, we introduce a new approach to this problem based on masking. Given an existing image-to-image network, our approach computes a mask such that the distance between the masked reconstructed image and the masked true image is guaranteed to be less than a specified threshold, with high probability. The mask thus identifies the more certain regions of the reconstructed image. Our approach is agnostic to the underlying image-to-image network, and only requires triples of the input (degraded), reconstructed and true images for training. Furthermore, our method is agnostic to the distance metric used. As a result, one can use L_p-style distances or perceptual distances like LPIPS, which contrasts with interval-based approaches to uncertainty. Our theoretical guarantees derive from a conformal calibration procedure. We evaluate our mask-based approach to uncertainty on image colorization, image completion, and super-resolution tasks, demonstrating high quality performance on each.

****

A Time-Consistency Curriculum for Learning from Instance-Dependent Noisy Labels
Songhua Wu,Tianyi Zhou,Yuxuan Du,Jun Yu,Bo Han,Tongliang Liu

Many machine learning algorithms are known to be fragile on simple instance-independent noisy labels. However, noisy labels in real-world data are more devastating since they are produced by more complicated mechanisms in an instance-dependent manner. In this paper, we target this practical challenge of \textit{Instance-Dependent Noisy Labels} by jointly training
(1) a model reversely engineering the noise generating mechanism, which produces an \textit{instance-dependent mapping} between the clean label posterior and the observed noisy label; and (2) a robust classifier that produces clean label posteriors. Compared to previous methods, the former model is novel and enables end-to-end learning of the latter directly from noisy labels. An extensive empirical study indicates that the time-consistency of data is critical to the success of training both models and motivates us to develop a curriculum selecting training data based on their dynamics on the two models' outputs over the course of training. We show that the curriculum-selected data provide both clean labels and high-quality input-output pairs for training the two models. Therefore, it leads to promising and robust classification performance even in notably challenging settings of instance-dependent noisy labels where many SoTA methods could easily fail. Extensive experimental comparisons and ablation studies further demonstrate the advantages and significance of the time-consistency curriculum in learning from instance-dependent noisy labels on multiple benchmark datasets.

****

Efficient Personalized Federated Learning via Sparse Model-Adaptation
Daoyuan Chen,Liuyi Yao,Dawei Gao,Bolin Ding,Yaliang Li

Federated Learning (FL) aims to train machine learning models for multiple clients without sharing their own private data. Due to the heterogeneity of clients' local data distribution, recent studies explore the personalized FL that learns and deploys distinct local models with the help of auxiliary global models. However, the clients can be heterogeneous in terms of not only local data distribution, but also their computation and communication resources. The capacity and efficiency of personalized models are restricted by the lowest-resource clients, leading to sub-optimal performance and limited practicality of personalized FL. To overcome these challenges, we propose a novel approach named pFedGate for efficient personalized FL by adaptively and efficiently learning sparse local models. With a lightweight trainable gating layer, pFedGate enables clients to reach their full potential in model capacity by generating different sparse models accounting for both the heterogeneous data distributions and resource constraints. Meanwhile, the computation and communication efficiency are both improved thanks to the adaptability between the model sparsity and clients' resources. Further, we theoretically show that the proposed pFedGate has superior complexity with guaranteed convergence and generalization error. Extensive experiments show that pFedGate achieves superior global accuracy, individual accuracy and efficiency simultaneously over state-of-the-art methods, by up to 4.53\% accuracy improvement and 12x smaller model size. We also demonstrate that pFedGate performs better than competitors in the novel clients participation and partial clients participation scenarios, and can learn meaningful sparse local models adapted to different data distributions.

****

Momentum Tracking: Momentum Acceleration for Decentralized Deep Learning on Heterogeneous Data
Yuki Takezawa,Han Bao,Kenta Niwa,Ryoma Sato,Makoto Yamada

SGD with momentum acceleration is one of the key components for improving the performance of neural networks. For decentralized learning, a straightforward approach using momentum acceleration is Distributed SGD (DSGD) with momentum acceleration (DSGDm). However, DSGDm performs worse than DSGD when the data distributions are statistically heterogeneous. Recently, several studies have addressed this issue and proposed methods with momentum acceleration that are more robust to data heterogeneity than DSGDm, although their convergence rates remain dependent on data heterogeneity and decrease when the data distributions are heterogeneous. In this study, we propose Momentum Tracking, which is a method with momentum acceleration whose convergence rate is proven to be independent of data heterogeneity. More specifically, we analyze the convergence rate of Momentum Tracking in the standard deep learning setting, where the objective function is non-convex and the stochastic gradient is used. Then, we identify that it is independent of data heterogeneity for any momentum coefficient $\beta \in [0, 1)$. Through image classification tasks, we demonstrate that Momentum Tracking is more robust to data heterogeneity than the existing decentralized learning methods with momentum acceleration and can consistently outperform these existing methods when the data distributions are heterogeneous.

****

Deep Graph-Level Orthogonal Hypersphere Compression for Anomaly Detection

Yunhe Zhang,Yan Sun,Jinyu Cai,Jicong Fan

Graph-level anomaly detection aims to identify abnormal samples of a set of graphs in an unsupervised manner. It is non-trivial to find a reasonable decision boundary between normal data and anomalous data without using any anomalous data in the training stage, especially for data in graphs. This paper first proposes a novel deep graph-level anomaly detection model, which learns the graph representation with maximum mutual information between substructure features and global structure features while exploring a hypersphere anomaly decision boundary. The deep orthogonal projection layer is adopted to keep the training data distribution consistent with the decision hypersphere thus avoiding erroneous evaluations. We further propose projecting the normal data into the interval region between two co-centered hyperspheres, which makes the normal data distribution more compact and effectively overcomes the issue of outliers falling close to the center of the hypersphere. The numerical and visualization results on a few graph datasets demonstrate the effectiveness and superiority of our methods in comparison to many baselines and state-of-the-art.

****

Gradient Deconfliction via Orthogonal Projections onto Subspaces For Multi-task Learning

Shijie Zhu,Hui Zhao,Pengjie Wang,Hongbo Deng,Jian Xu,Bo Zheng

Although multi-task learning (MTL) has been a preferred approach and successfully applied in many real-world scenarios, MTL models are not guaranteed to outperform single-task models on all tasks mainly due to the negative effects of conflicting gradients among the tasks. In this paper, we fully examine the influence of conflicting gradients and further emphasize the importance and advantages of achieving non-conflicting gradients which allows simple but effective trade-off strategies among the tasks with stable performance. Base on our findings, we propose the Gradient Deconfliction via Orthogonal Projections onto Subspaces (GradOPS) spanned by other task-specific gradients. Our method not only solves all conflicts among the tasks, but can also effectively search for diverse solutions towards different trade-off preferences among the tasks. Theoretical analysis on convergence is provided, and performance of our algorithm is fully testified on multiple benchmarks in various domains. Results demonstrate that our method can effectively find multiple state-of-the-art solutions with different trade-off strategies among the tasks on multiple datasets.

****

Pareto Optimization for Active Learning under Out-of-Distribution Data Scenarios

Xueying Zhan,Zeyu Dai,Qingzhong Wang,Haoyi Xiong,Dejing Dou,Qing Li,Antoni B. Chan

Pool-based Active Learning (AL) has achieved great success in minimizing labeling costs by sequentially selecting the most informative unlabeled samples from a large unlabeled data pool and querying their labels from oracle/annotators. However, existing AL sampling schemes might not work well under out-of-distribution (OOD) data scenarios, where the unlabeled data pool contains data samples that do not belong to the pre-defined categories of the target task. Achieving good AL performance under OOD data scenarios is a challenging task due to the natural conflict between AL sampling strategies and OOD sample detection -- both more inf ormative in-distribution (ID) data and OOD data in unlabeled data pool may be assigned high informativeness scores (e.g., high entropy) during AL processes. In this paper, we propose a sampling scheme, Monte-Carlo Pareto Optimization for Active Learning (POAL), which selects optimal subsets of unlabeled samples with \emph{fixed batch size} from the unlabeled data pool. We cast the AL sampling task as a multi-objective optimization problem and utilize Pareto optimization based on two conflicting objectives: (1) the typical AL sampling scheme (e.g., maximum entropy), and (2) the confidence of not being an OOD data sample. Experimental results show the effectiveness of our POAL on classical Machine Learning (ML) and Deep Learning (DL) tasks.

****

Self-Consistent Learning: Cooperation between Generators and Discriminators
Tong Wu,Hao Wang,Zhongshen Zeng,Wei Wang,Hai-Tao Zheng,Jiaxing Zhang

Using generated data to improve the performance of downstream discriminative models has recently gained popularity due to the great development of pre-trained language models. In most previous studies, generative models and discriminative models are trained separately and thus could not adapt to any changes in each other. As a result, the generated samples can easily deviate from the real data distribution, while the improvement of the discriminative model quickly reaches saturation. Generative adversarial networks (GANs) train generative models via an adversarial process with discriminative models to achieve joint training. However, the training of standard GANs is notoriously unstable and often falls short of convergence. In this paper, to address these issues, we propose a $\textit{self-consistent learning}$ framework, in which a discriminator and a generator are cooperatively trained in a closed-loop form. The discriminator and the generator enhance each other during multiple rounds of alternating training until a scoring consensus is reached. This framework proves to be easy to train and free from instabilities such as mode collapse and non-convergence. Extensive experiments on sentence semantic matching demonstrate the effectiveness of the proposed framework: the discriminator achieves 10+ AP of improvement on the zero-shot setting and new state-of-the-art performance on the full-data setting.

****

Learning Dynamical Characteristics with Neural Operators for Data Assimilation
Yi Xiao,Wei Xue

Data assimilation refers to a group of algorithms that combines numerical models with observations to obtain an optimal estimation of the system's states. In areas like earth science, numerical models are usually formulated by differential equations, also known as the prior dynamics. It is a great challenge for neural networks to properly exploit the dynamical characteristics for data assimilation, because first, it is difficult to represent complicated dynamical characteristics in neural networks, and second, the dynamics are likely to be biased. The state-of-the-art neural networks borrow from the traditional method to introduce dynamical characteristics by optimizing the 4D-Var objective function in which the dynamics are inherently quantified, but the iterative optimization process leads to high computational cost. In this paper, we develop a novel deep learning framework with neural operators for data assimilation. The key novelty of our proposed approach is that we design a so-called flow operator through self-supervised learning to explicitly learn dynamical characteristics for reconstructed states. Numerical experiments on the Lorenz-63 and Lorenz-96 systems, which are the standard benchmarks for data assimilation performance evaluation, show that the proposed method is at least three times faster than state-of-the-art neural networks, and reduces the dynamic loss by two orders of magnitude. It is also demonstrated that our method is well-adapted to biases in the prior dynamics.

****

Lost Domain Generalization Is a Natural Consequence of Lack of Training Domains
Yimu Wang,Hongyang Zhang

We show a hardness result for the number of training domains required to achieve a small population error in the test domain. Although many domain generalization algorithms have been developed under various domain-invariance assumptions, there is significant evidence to indicate that out-of-distribution (o.o.d.) test a ccuracy of state-of-the-art o.o.d. algorithms is on par with empirical risk minimization and random guess on the domain generalization benchmarks such as DomainBed. In this work, we analyze its cause and attribute the lost domain generalization to the lack of training domains. We show that, in a minimax lower bound fashion, \emph{any} learning algorithm that outputs a classifier with an ϵ excess error to the Bayes optimal classifier requires at least $\mathrm{poly}(1/\epsilon)$ number of training domains, even though the number of training data sampled from each training domain is large. Experiments on the DomainBed benchmark demonstrate that o.o.d. test accuracy is monotonically increasing as the number of training domains increases. Our result sheds light on the intrinsic hardness of domain generalization and suggests benchmarking o.o.d. algorithms by the datasets with a sufficient number of training domains.

****

Graph Neural Networks for Link Prediction with Subgraph Sketching

Benjamin Paul Chamberlain,Sergey Shirobokov,Emanuele Rossi,Fabrizio Frasca,Thomas Markovich,Nils Yannick Hammerla,Michael M. Bronstein,Max Hansmire

Many Graph Neural Networks (GNNs) perform poorly compared to simple heuristics on Link Prediction (LP) tasks. This is due to limitations in expressive power such as the inability to count triangles (the backbone of most LP heuristics) and because they can not distinguish automorphic nodes (those having identical structural roles). Both expressiveness issues can be alleviated by learning link (rather than node) representations and incorporating structural features such as triangle counts. Since explicit link representations are often prohibitively expensive, recent works resorted to subgraph-based methods, which have achieved state-of-the-art performance for LP, but suffer from poor efficiency due to high levels of redundancy between subgraphs. We analyze the components of subgraph GNN (SGNN) methods for link prediction. Based on our analysis, we propose a novel full-graph GNN called ELPH (Efficient Link Prediction with Hashing) that passes subgraph sketches as messages to approximate the key components of SGNNs without explicit subgraph construction. ELPH is provably more expressive than Message Passing GNNs (MPNNs). It outperforms existing SGNN models on many standard LP benchmarks while being orders of magnitude faster. However, it shares the common GNN limitation that it is only efficient when the dataset fits in GPU memory. Accordingly, we develop a highly scalable model, called BUDDY, which uses feature precomputation to circumvent this limitation without sacrificing predictive performance. Our experiments show that BUDDY also outperforms SGNNs on standard LP benchmarks while being highly scalable and faster than ELPH.

****

Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning

Zhen Wang,Rameswar Panda,Leonid Karlinsky,Rogerio Feris,Huan Sun,Yoon Kim

Prompt tuning, in which a base pretrained model is adapted to each task via conditioning on learned prompt vectors, has emerged as a promising approach for efficiently adapting large language models to multiple downstream tasks. However, existing methods typically learn soft prompt vectors from scratch, and it has not been clear how to exploit the rich cross-task knowledge with prompt vectors in a multitask learning setting. We propose multitask prompt tuning (MPT), which first learns a single transferable prompt by distilling knowledge from multiple task-specific source prompts. We then learn multiplicative low rank updates to this shared prompt to efficiently adapt it to each downstream target task. Extensive experiments on 23 NLP datasets demonstrate that our proposed approach outperforms the state-of-the-art methods, including the full finetuning baseline in some cases, despite only tuning 0.035% as many task-specific parameters.

****

Personalized Federated Hypernetworks for Privacy Preservation in Multi-Task Reinforcement Learning

Doseok Jang,Larry Yan,Lucas Spangher,Selvaprabu Nadarajah,Costas Spanos

Multi-Agent Reinforcement Learning currently focuses on implementations where all data and training can be centralized to one machine. But what if local agents are split across multiple tasks, and need to keep data private between each? We develop the first application of Personalized Federated Hypernetworks (PFH) to R

einforcement Learning (RL). We then present a novel application of PFH to few-shot transfer, and demonstrate significant initial increases in learning. PFH has never been demonstrated beyond supervised learning benchmarks, so we apply PFH to an important domain: RL price-setting for energy demand response. We consider a general case across where agents are split across multiple microgrids, wherein energy consumption data must be kept private within each microgrid. Together, our work explores how the fields of personalized federated learning and RL can come together to make learning efficient across multiple tasks while keeping data secure.
****
CBLab: Scalable Traffic Simulation with Enriched Data Supporting
Chumeng Liang,Zherui Huang,Yicheng Liu,Zhanyu Liu,Guanjie Zheng,Hanyuan Shi,Yuhao Du,FULIANG LI,Zhenhui Li
Traffic simulation provides interactive data for the optimization of traffic policies. However, existing traffic simulators are limited by their lack of scalability and shortage in input data, which prevents them from generating interactive data from traffic simulation in the scenarios of real large-scale city road networks.

In this paper, we present \textbf{C}ity \textbf{B}rain \textbf{Lab}, a toolkit for scalable traffic simulation. CBLab is consist of three components: CBEngine, CBData, and CBScenario. CBEngine is a highly efficient simulator supporting large-scale traffic simulation. CBData includes a traffic dataset with road network data of 100 cities all around the world. We also develop a pipeline to conduct a one-click transformation from raw road networks to input data of our traffic simulation. Combining CBEngine and CBData allows researchers to run scalable traffic simulations in the road network of real large-scale cities. Based on that, CBScenario implements an interactive environment and several baseline methods for two scenarios of traffic policies respectively, with which traffic policies adaptable for large-scale urban traffic can be trained and tuned. To the best of our knowledge, CBLab is the first infrastructure supporting traffic policy optimization in large-scale urban scenarios. The code is available on GitHub:~\url{https://github.com/CityBrainLab/CityBrainLab.git}.
****
Personalized Decentralized Bilevel Optimization over Stochastic and Directed Networks
Naoyuki Terashita,Satoshi Hara
While personalization in distributed learning has been extensively studied, existing approaches employ dedicated algorithms to optimize their specific type of parameters (e.g., client clusters or model interpolation weights), making it difficult to simultaneously optimize different types of parameters to yield better performance.
Moreover, their algorithms require centralized or static undirected communication networks, which can be vulnerable to center-point failures or deadlocks.
This study proposes optimizing various types of parameters using a single algorithm that runs on more practical communication environments.
First, we propose a gradient-based bilevel optimization that reduces most personalization approaches to the optimization of client-wise hyperparameters.
Second, we propose a decentralized algorithm to estimate gradients with respect to the hyperparameters, which can run even on stochastic and directed communication networks.
Our empirical results demonstrated that the gradient-based bilevel optimization enabled combining existing personalization approaches which led to state-of-the-art performance, confirming it can perform on multiple simulated communication environments including a stochastic and directed network.
****
ContextSpeech: Expressive and Efficient Text-to-Speech for Paragraph Reading
Yujia Xiao,Shaofei Zhang,Xi Wang,Xu Tan,Lei He,Frank K. Soong,sheng zhao
Although Text-to-Speech (TTS) has made rapid progress in speech quality at sentence level, it still faces a lot of challenges in paragraph / long-form reading.

Synthesizing sentence by sentence in a paragraph and then concatenating them together will cause inconsistent issues that affect paragraph-level expressiveness. While directly modelling all the sentences in a paragraph will incur large computation / memory cost. In this paper, we develop a TTS system called ContextSpeech, which models the contextual information in a paragraph for coherence and expressiveness without largely increasing the computation or memory cost. On the one hand, we introduce a memory-cached recurrence mechanism to let the current sentence see more history information both on the text and speech sides. On the other hand, we construct text-based semantic information in a hierarchical structure, which can broaden the horizon and incorporate the future information. Additionally, we use a linearized self-attention with compatible relative-position encoding to reduce the computation / memory cost. Experiments show that ContextSpeech significantly improves the paragraph-level voice quality and prosody expressiveness in terms of both subjective and objective evaluation metrics. Furthermore, ContextSpeech achieves better model efficiency in both training and inference stage.

****

Learning Object Affordance with Contact and Grasp Generation

Haoming Li,Xinzhuo Lin,Yang Zhou,Xiang Li,Jiming Chen,Qi Ye

Understanding object affordance can help in designing better and more robust robotic grasping. Existing work in the computer vision community formulates the object affordance understanding as a grasping pose generation problem, which treats the problem as a black box by learning a mapping between objects and the distributions of possible grasping poses for the objects. On the other hand, in the robotics community, estimating object affordance represented by contact maps is of the most importance as localizing the positions of the possible affordance can help the planning of grasping actions. In this paper, we propose to formulate the object affordance understanding as both contacts and grasp poses generation. we factorize the learning task into two sequential stages, rather than the black-box strategy: (1) we first reason the contact maps by allowing multi-modal contact generation; (2) assuming that grasping poses are fully constrained given contact maps, we learn a one-to-one mapping from the contact maps to the grasping poses. Further, we propose a penetration-aware partial optimization from the intermediate contacts. It combines local and global optimization for the refinement of the partial poses of the generated grasps exhibiting penetration. Extensive validations on two public datasets show our method outperforms state-of-the-art methods regarding grasp generation on various metrics.

****

Deep Graph-Level Clustering Using Pseudo-Label-Guided Mutual Information Maximization Network

Jinyu Cai,Yi Han,Wenzhong Guo,Jicong Fan

In this work, we study the problem of partitioning a set of graphs into different groups such that the graphs in the same group are similar while the graphs in different groups are dissimilar. This problem was rarely studied previously, although there have been a lot of work on node clustering and graph classification. The problem is challenging because it is difficult to measure the similarity or distance between graphs. One feasible approach is using graph kernels to compute a similarity matrix for the graphs and then performing spectral clustering, but the effectiveness of existing graph kernels in measuring the similarity between graphs is very limited. To solve the problem, we propose a novel method called Deep Graph-Level Clustering (DGLC). DGLC utilizes a graph isomorphism network to learn graph-level representations by maximizing the mutual information between the representations of entire graphs and substructures, under the regularization of a clustering module that ensures discriminative representations via pseudo labels. DGLC achieves graph-level representation learning and graph-level clustering in an end-to-end manner. The experimental results on six benchmark datasets of graphs show that our DGLC has state-of-the-art performance in comparison to many baselines.

****

Deep Generative Model based Rate-Distortion for Image Downscaling Assessment

Yuanbang Liang,Bhavesh Garg,Paul L Rosin,Yipeng Qin

In this paper, we propose a novel measure, namely Image Downscaling Assessment by Rate-Distortion (IDA-RD), to quantitatively evaluate image downscaling algorithms. In contrast to image-based methods that measure the quality of downscaled images, ours is process-based that draws ideas from the rate-distortion theory to measure the distortion incurred during downscaling. Our main idea is that downscaling and super-resolution (SR) can be viewed as the encoding and decoding processes in the rate-distortion model, respectively, and that a downscaling algorithm that preserves more details in the resulting low-resolution (LR) images should lead to less distorted high-resolution (HR) images in SR. In other words, the distortion should increase as the downscaling algorithm deteriorates. However, it is non-trivial to measure this distortion as it requires the SR algorithm to be blind and stochastic. Our key insight is that such requirements can be met by recent SR algorithms based on deep generative models that can find all matching HR images for a given LR image on their learned image manifolds. Empirically, we first validate our IDA-RD measure with synthetic downscaling algorithms which simulate distortions by adding various types and levels of degradations to the downscaled images. We then test our measure on traditional downscaling algorithms such as bicubic, bilinear, nearest neighbor interpolation as well as state-of-the-art downscaling algorithms such as DPID, L0-regularized downscaling, and Perceptual downscaling. Experimental results show the effectiveness of our IDA-RD in evaluating image downscaling algorithms.

****

Selective Classifier Ensemble

Qiang Ding,Yixuan Cao,Ping Luo

Selective classification allows a machine learning model to abstain on some hard inputs and thus improve the safety of its predictions. In this paper, we study the ensemble of selective classifiers, i.e. selective classifier ensemble, which combines several weak selective classifiers to obtain a more powerful model. We prove that under some assumptions, the ensemble has a lower selective risk than the individual model under a range of coverage. This is nontrivial since the selective risk is a non-convex function of the model prediction. The assumptions and the theoretical result are supported by systematic experiments on both computer vision and natural language processing tasks. A surprising empirical result is that a simple selective classifier ensemble, namely, the ensemble model with maximum probability as confidence, is the state-of-the-art selective classifier. For instance, on CIFAR-10, using the same VGG-16 backbone model, this ensemble reduces the AURC (Area Under Risk-Coverage Curve) by about 24%, relative to the previous state-of-the-art method.

****

Better Generative Replay for Continual Federated Learning

Daiqing Qi,Handong Zhao,Sheng Li

Federated Learning (FL) aims to develop a centralized server that learns from distributed clients via communications without accessing the clients’ local data. However, existing works mainly focus on federated learning in a single task scenario with static data. In this paper, we introduce the continual federated learning (CFL) problem, where clients incrementally learn new tasks and history data can- not be stored due to certain reasons, such as limited storage and data retention policy 1. Generative replay (GR) based methods are effective for continual learning without storing history data. However, we fail when trying to intuitively adapt GR models for this setting. By analyzing the behaviors of clients during training, we find the unstable training process caused by distributed training on non-IID data leads to a notable performance degradation. To address this problem, we propose our FedCIL model with two simple but effective solutions: 1. model consolidation and 2. consistency enforcement. Experimental results on multiple benchmark datasets demonstrate that our method significantly outperforms baselines. Code is available at: https://github.com/daiqing98/FedCIL.

****

Enhancing the Transferability of Adversarial Examples via a Few Queries and Fuzzy Domain Eliminating

Xiangyuan Yang,Jie Lin,hanlin Zhang,Xinyu Yang,Peng Zhao

Due to the vulnerability of deep neural networks, the black-box attack has drawn great attention from the community. Though transferable priors decrease the query number of the black-box query attacks in recent efforts, the average number of queries is still larger than 100, which is easily affected by the number of queries limit policy. In this work, we propose a novel method called query prior-based method to enhance the attack transferability of the family of fast gradient sign methods by using a few queries. Specifically, for the untargeted attack, we find that the successful attacked adversarial examples prefer to be classified as the wrong categories with higher probability by the victim model. Therefore, the weighted augmented cross-entropy loss is proposed to reduce the gradient angle between the surrogate model and the victim model for enhancing the transferability of the adversarial examples. In addition, the fuzzy domain eliminating technique is proposed to avoid the generated adversarial examples getting stuck in the local optimum. Specifically, we define the fuzzy domain of the input example x in the ϵ-ball of x. Then, temperature scaling and fuzzy scaling are utilized to eliminate the fuzzy domain for enhancing the transferability of the generated adversarial examples. Theoretical analysis and extensive experiments demonstrate that our method could significantly improve the transferability of gradient-based adversarial attacks on CIFAR10/100 and ImageNet and outperform the black-box query attack with the same few queries.

****

Label-distribution-agnostic Ensemble Learning on Federated Long-tailed Data

Yaopei Zeng,Lei Liu,Baoyuan Wu,ShaoGuo Liu,Li Liu

Federated Learning (FL) is a distributed machine learning paradigm that enables devices to collaboratively train a shared model. However, the long-tailed distribution in nature deteriorates the performance of the global model, which is difficult to address due to data heterogeneity, e.g., local clients may exhibit diverse imbalanced class distributions. Moreover, existing re-balance strategies generally utilize label distribution as the class prior, which may conflict with the privacy requirement of FL. To this end, we propose a Label-Distribution-Agnostic Ensemble (LDAE) learning framework to integrate heterogeneous data distributions using multiple experts, which targets to optimize a balanced global objective under privacy protection. In particular, we derive a privacy-preserving proxy from the model updates of clients to guide the grouping and updating of multiple experts. Knowledge from clients can be aggregated via implicit interactions among different expert groups. We theoretically and experimentally demonstrate that (1) there is a global objective gap between global and local re-balance strategies\footnote{The local re-balance strategy means that each client utilizes re-balance methods based on the local label distribution, while the global re-balance strategy applies re-balance methods using global label distribution as the class-wise prior.} and (2) with protecting data privacy, the proxy can be used as an alternative to label distribution for existing class prior based re-balance strategies. Extensive experiments on long-tailed decentralized datasets demonstrate the effectiveness of our method, showing superior performance to state-of-the-art methods.

****

Generative Modelling with Inverse Heat Dissipation

Severi Rissanen,Markus Heinonen,Arno Solin

While diffusion models have shown great success in image generation, their noise-inverting generative process does not explicitly consider the structure of images, such as their inherent multi-scale nature. Inspired by diffusion models and the empirical success of coarse-to-fine modelling, we propose a new diffusion-like model that generates images through stochastically reversing the heat equation, a PDE that locally erases fine-scale information when run over the 2D plane of the image. We interpret the solution of the forward heat equation with constant additive noise as a variational approximation in the diffusion latent variable model. Our new model shows emergent qualitative properties not seen in standard diffusion models, such as disentanglement of overall colour and shape in images. Spectral analysis on natural images highlights connections to diffusion models and reveals an implicit coarse-to-fine inductive bias in them.

****

Self-supervision through Random Segments with Autoregressive Coding (RandSAC)

Tianyu Hua,Yonglong Tian,Sucheng Ren,Michalis Raptis,Hang Zhao,Leonid Sigal

Inspired by the success of self-supervised autoregressive representation learning in natural language (GPT and its variants), and advances in recent visual architecture design with Vision Transformers (ViTs), in this paper, we explore the effects various design choices have on the success of applying such training strategies for visual feature learning. Specifically, we introduce a novel strategy that we call Random Segments with Autoregressive Coding (RandSAC). In RandSAC, we group patch representations (image tokens) into hierarchically arranged segments; within each segment, tokens are predicted in parallel, similar to BERT, while across segment predictions are sequential, similar to GPT. We illustrate that randomized serialization of the segments significantly improves the performance and results in distribution over spatially-long (across-segments) and -short (within-segment) predictions which are effective for feature learning. We illustrate the pertinence of these design choices and explore alternatives on a number of datasets (e.g., CIFAR10, ImageNet). While our pre-training strategy works with vanilla Transformer, we also propose a conceptually simple, but highly effective, addition to the decoder that allows learnable skip-connections to encoder feature layers, which further improves the performance.

****

Rarity Score : A New Metric to Evaluate the Uncommonness of Synthesized Images

Jiyeon Han,Hwanil Choi,Yunjey Choi,Junho Kim,Jung-Woo Ha,Jaesik Choi

Evaluation metrics in image synthesis play a key role to measure performances of generative models. However, most metrics mainly focus on image fidelity. Existing diversity metrics are derived by comparing distributions, and thus they cannot quantify the diversity or rarity degree of each generated image. In this work, we propose a new evaluation metric, called `rarity score', to measure both image-wise uncommonness and model-wise diversified generation performance.

We first show empirical observation that typical samples are close to each other and distinctive samples are far from each other in nearest-neighbor distances on latent spaces represented by feature extractor networks such as VGG16. We then show that one can effectively filter typical or distinctive samples with the proposed metric. We also use our metric to demonstrate that the extent to which different generative models produce rare images can be effectively compared. Further, our metric can be used to compare rarities between datasets that share the same concept such as CelebA-HQ and FFHQ. Finally, we analyze the use of metrics in different designs of feature extractors to better understand the relationship between feature spaces and resulting high-rarity images. Code will be publicly available for the research community.

****

Benchmarking Approximate k-Nearest Neighbour Search for Big High Dimensional Dynamic Data

Ben Harwood,Amir Dezfouli,Iadine Chades

Approximate k-Nearest Neighbour (ANN) methods are commonly used for mining information from big high-dimensional datasets. For each application the high-level dataset properties and run-time requirements determine which method will provide the most suitable tradeoffs. However, due to a significant lack of comprehensive benchmarking, judicious method selection is not currently possible for ANN applications that involve frequent online changes to datasets. Here we address this issue by building upon existing benchmarks for static search problems to provide a new benchmarking framework for big high dimensional dynamic data. We apply our framework to dynamic scenarios modelled after common real world applications. In all cases we are able to identify a suitable recall-runtime tradeoff to improve upon a worst-case exhaustive search. Our framework provides a flexible solution to accelerate future ANN research and enable researchers in other online data-rich domains to find suitable methods for handling their ANN searches.

****

Semi-Supervised Offline Reinforcement Learning with Action-Free Trajectories

Qinqing Zheng,Mikael Henaff,Brandon Amos,Aditya Grover

Natural agents can effectively learn from multiple data sources that differ in size, quality, and types of measurements. We study this heterogeneity in the context of offline reinforcement learning (RL) by introducing a new, practically motivated semi-supervised setting. Here, an agent has access to two sets of trajectories: labelled trajectories containing state, action, reward triplets at every timestep, along with unlabelled trajectories that contain only state and reward information. For this setting, we develop a simple meta-algorithmic pipeline that learns an inverse-dynamics model on the labelled data to obtain proxy-labels for the unlabelled data, followed by the use of any offline RL algorithm on the true and proxy-labelled trajectories. Empirically, we find this simple pipeline to be highly successful --- on several D4RL benchmarks~\cite{fu2020d4rl}, certain offline RL algorithms can match the performance of variants trained on a fully labeled dataset even when we label only 10\% trajectories from the low return regime. Finally, we perform a large-scale controlled empirical study investigating the interplay of data-centric properties of the labelled and unlabelled datasets, with algorithmic design choices (e.g., inverse dynamics, offline RL algorithm) to identify general trends and best practices for training RL agents on semi-supervised offline datasets.

****

E-Forcing: Improving Autoregressive Models by Treating it as an Energy-Based One

Yezhen Wang,Tong Che,Bo Li,Kaitao Song,Hengzhi Pei,Yoshua Bengio,Dongsheng Li

Autoregressive generative models are commonly used to solve tasks involving sequential data. They have, however, been plagued by a slew of inherent flaws due to the intrinsic characteristics of chain-style conditional modeling (e.g., exposure bias or lack of long-range coherence), severely limiting their ability to model distributions properly. In this paper, we propose a unique method termed E-Forcing for training autoregressive generative models that takes advantage of a well-designed energy-based learning objective. By leveraging the extra degree of freedom of the softmax operation, we are allowed to make the autoregressive model itself an energy-based model for measuring the likelihood of input without introducing any extra parameters. Furthermore, we show that with the help of E-Forcing, we can alleviate the above flaws for autoregressive models. Extensive empirical results, covering numerous benchmarks demonstrate the effectiveness of the proposed approach.

****

Masked Vector Quantization

David D Nguyen,David Liebowitz,Surya Nepal,Salil S. Kanhere

Generative models with discrete latent representations have recently demonstrated an impressive ability to learn complex high-dimensional data distributions. However, their performance relies on a long sequence of tokens per instance and a large number of codebook entries, resulting in long sampling times and considerable computation to fit the categorical posterior.

To address these issues, we propose the Masked Vector Quantization (MVQ) framework which increases the representational capacity of each code vector by learning mask configurations via a stochastic winner-takes-all training regime called Multiple Hypotheses Dropout (MH-Dropout). On ImageNet 64$\times$64, reduces FID in existing vector quantization architectures by up to 68% at 2 tokens per instance and 57% at 5 tokens. These improvements widen as codebook entries is reduced and allows for $7\textup{-}45\times$ speed-up in token sampling during inference. As an additional benefit, we find that smaller latent spaces lead to MVQ identifying transferable visual representations where multiple can be smoothly combined.

****

On the Importance of the Policy Structure in Offline Reinforcement Learning

Takayuki Osa,Akinobu Hayashi,Pranav Deo,Naoki Morihira,Takahide Yoshiike

Offline reinforcement learning (RL) has attracted a great deal of attention recently as an approach to utilizing past experience to learn a policy. Recent studies have reported the challenges of offline RL, such as estimating the values of actions that are out of the data distribution. To mitigate the issues of offline

RL, we propose an algorithm that leverages a mixture of deterministic policies. With our framework, the state-action space is divided by learning discrete latent variables, and sub-policies corresponding to each region are trained. The proposed algorithm, which we call Value-Weighted Variational Auto-Encoder (V2AE), is derived by considering the variational lower bound of the offline RL objective function. The aim of this work is to shed lights on the importance on the policy structure in offline RL. We show empirically that the use of the proposed mixture policy can reduce the accumulation of the approximation error in offline RL, which was reported in previous studies. Experimental results also indicate that introducing the policy structure improves the performance on tasks with D4RL benchmarking datasets.
****

Bandit Learning in Many-to-one Matching Markets with Uniqueness Conditions
Liya Guo,Zilong Wang,Shuai Li
An emerging line of research is dedicated to the problem of one-to-one matching markets with bandits, where the preference of one side is unknown and thus we need to match while learning the preference through multiple rounds of interaction. However, in many real-world applications such as online recruitment platform for short-term workers, one side of the market can select more than one participant from the other side, which motivates the study of the many-to-one matching problem. Moreover, the existence of a unique stable matching is crucial to the competitive equilibrium of the market. In this paper, we first introduce a more general new \textit{$\tilde{\alpha}$}-condition to guarantee the uniqueness of stable matching in many-to-one matching problems, which generalizes some established uniqueness conditions such as \textit{SPC} and \textit{Serial Dictatorship}, and recovers the known α-condition if the problem is reduced to one-to-one matching. Under this new condition, we design an MO-UCB-D4 algorithm with $O\left(\frac{NK\log(T)}{\Delta^2}\right)$ regret bound, where T is the time horizon, N is the number of agents, K is the number of arms, and Δ is the minimum reward gap. Extensive experiments show that our algorithm achieves uniform good performances under different uniqueness conditions.
****

Transformer-Patcher: One Mistake Worth One Neuron
Zeyu Huang,Yikang Shen,Xiaofeng Zhang,Jie Zhou,Wenge Rong,Zhang Xiong
Large Transformer-based Pretrained Language Models (PLMs) dominate almost all Natural Language Processing (NLP) tasks. Nevertheless, they still make mistakes from time to time. For a model deployed in an industrial environment, fixing these mistakes quickly and robustly is vital to improve user experiences. Previous works formalize such problems as Model Editing (ME) and mostly focus on fixing one mistake. However, the one-mistake-fixing scenario is not an accurate abstraction of the real-world challenge. In the deployment of AI services, there are ever-emerging mistakes, and the same mistake may recur if not corrected in time. Thus a preferable solution is to rectify the mistakes as soon as they appear nonstop. Therefore, we extend the existing ME into the Sequential Model Editing (SME) to help develop more practical editing methods. Our study shows that current ME methods either fail to make a sequence of edits or to remember previous edits. We then introduce Transformer-Patcher, a novel model editor that can shift the behavior of transformer-based models by simply adding and training a few neurons in the last Feed-Forward Network layer. Experimental results on both classification and generation tasks show that Transformer-Patcher can successively correct up to thousands of errors (Reliability) and generalize to their equivalent inputs (Generality) while retaining the model’s accuracy on irrelevant inputs (Locality). Our method outperforms previous fine-tuning and HyperNetwork-based methods and achieves state-of-the-art performance for Sequential Model Editing (SME).
****

Corrupted Image Modeling for Self-Supervised Visual Pre-Training
Yuxin Fang,Li Dong,Hangbo Bao,Xinggang Wang,Furu Wei
We introduce Corrupted Image Modeling (CIM) for self-supervised visual pre-training. CIM uses an auxiliary generator with a small trainable BEiT to corrupt the input image instead of using artificial [MASK] tokens, where some patches are ra

ndomly selected and replaced with plausible alternatives sampled from the BEiT output distribution. Given this corrupted image, an enhancer network learns to either recover all the original image pixels, or predict whether each visual token is replaced by a generator sample or not. The generator and the enhancer are simultaneously trained and synergistically updated. After pre-training, the enhancer can be used as a high-capacity visual encoder for downstream tasks. CIM is a general and flexible visual pre-training framework that is suitable for various network architectures. For the first time, CIM demonstrates that both ViT and CNN can learn rich visual representations using a unified, non-Siamese framework. Experimental results show that our approach achieves compelling results in vision benchmarks, such as ImageNet classification and ADE20K semantic segmentation.
****

Semi-Implicit Variational Inference via Score Matching
Longlin Yu,Cheng Zhang

Semi-implicit variational inference (SIVI) greatly enriches the expressiveness of variational families by considering implicit variational distributions defined in a hierarchical manner. However, due to the intractable densities of variational distributions, current SIVI approaches often use surrogate evidence lower bounds (ELBOs) or employ expensive inner-loop MCMC runs for unbiased ELBOs for training. In this paper, we propose SIVI-SM, a new method for SIVI based on an alternative training objective via score matching. Leveraging the hierarchical structure of semi-implicit variational families, the score matching objective allows a minimax formulation where the intractable variational densities can be naturally handled with denoising score matching. We show that SIVI-SM closely matches the accuracy of MCMC and outperforms ELBO-based SIVI methods in a variety of Bayesian inference tasks.
****

Sharper Bounds for Uniformly Stable Algorithms with Stationary Mixing Process
Shi Fu,Yunwen Lei,Qiong Cao,Xinmei Tian,Dacheng Tao

Generalization analysis of learning algorithms often builds on a critical assumption that training examples are independently and identically distributed, which is often violated in practical problems such as time series prediction. In this paper, we use algorithmic stability to study the generalization performance of learning algorithms with ψ-mixing data, where the dependency between observations weakens over time. We show uniformly stable algorithms guarantee high-probability generalization bounds of the order $O(1/\sqrt{n})$ (within a logarithmic factor), where n is the sample size. We apply our general result to specific algorithms including regularization schemes, stochastic gradient descent and localized iterative regularization, and develop excess population risk bounds for learning with ψ-mixing data. Our analysis builds on a novel moment bound for weakly-dependent random variables on a φ-mixing sequence and a novel error decomposition of generalization error.
****

Few-Shot Anomaly Detection on Industrial Images through Contrastive Fine-Tuning
Jingyi Liao,Xun Xu,Manh Cuong Nguyen,Chuan-Sheng Foo

Detecting abnormal products through imagery data is essential to the quality control in manufacturing. Existing approaches towards anomaly detection~(AD) often rely on substantial amount of anomaly-free samples to train representation and density models. Nevertheless, large anomaly-free datasets may not always be available before inference stage and this requires building an anomaly detection framework with only a handful of normal samples, a.k.a. few-shot anomaly detection (FSAD). We propose two techniques to address the challenges in FSAD. First, we employ a model pretrained on large source dataset to initialize model weights. To ameliorate the covariate shift between source and target domains, we adopt contrastive training on the few-shot target domain data. Second, to encourage learning representations suitable for downstream AD, we further incorporate cross-instance pairs to increase tightness within normal sample cluster and better separation between normal and synthesized negative samples. Extensive evaluations on six few-shot anomaly detection benchmarks demonstrate the effectiveness of the proposed method.

****

Rate-Distortion Optimized Post-Training Quantization for Learned Image Compression

Junqi Shi,Ming Lu,fangdong chen,Shiliang Pu,Zhan Ma

Quantizing floating-point neural network to its fixed-point representation is crucial for Learned Image Compression (LIC) because it ensures the decoding consistency for interoperability and reduces space-time complexity for implementation. Existing solutions often have to retrain the network for model quantization which is time consuming and impractical. This work suggests the use of Post-Training Quantization (PTQ) to directly process pretrained, off-the-shelf LIC models. We theoretically prove that minimizing the mean squared error (MSE) in PTQ is sub-optimal for compression task and thus develop a novel Rate-Distortion (R-D) Optimized PTQ (RDO-PTQ) to best retain the compression performance. Such RDO-PTQ just needs to compress few images (e.g., 10) to optimize the transformation of weight, bias, and activation of underlying LIC model from its native 32-bit floating-point (FP32) format to 8-bit fixed-point (INT8) precision for fixed-point inference onwards. Experiments reveal outstanding efficiency of the proposed method on different LICs, showing the closest coding performance to their floating-point counterparts. And, our method is a lightweight and plug-and-play approach without any need of model retraining which is attractive to practitioners.

****

Benign Overfitting in Classification: Provably Counter Label Noise with Larger Models

Kaiyue Wen,Jiaye Teng,Jingzhao Zhang

Studies on benign overfitting provide insights for the success of overparameterized deep learning models. In this work, we examine whether overfitting is truly benign in real-world classification tasks. We start with the observation that a ResNet model overfits benignly on Cifar10 but not benignly on ImageNet. To understand why benign overfitting fails in the ImageNet experiment, we theoretically analyze benign overfitting under a more restrictive setup where the number of parameters is not significantly larger than the number of data points. Under this mild overparameterization setup, our analysis identifies a phase change: unlike in the previous heavy overparameterization settings, benign overfitting can now fail in the presence of label noise. Our analysis explains our empirical observations, and is validated by a set of control experiments with ResNets. Our work highlights the importance of understanding implicit bias in underfitting regimes as a future direction.

****

Predictive Inference with Feature Conformal Prediction

Jiaye Teng,Chuan Wen,Dinghuai Zhang,Yoshua Bengio,Yang Gao,Yang Yuan

Conformal prediction is a distribution-free technique for establishing valid prediction intervals. Although conventionally people conduct conformal prediction in the output space, this is not the only possibility. In this paper, we propose feature conformal prediction, which extends the scope of conformal prediction to semantic feature spaces by leveraging the inductive bias of deep representation learning. From a theoretical perspective, we demonstrate that feature conformal prediction provably outperforms regular conformal prediction under mild assumptions. Our approach could be combined with not only vanilla conformal prediction, but also other adaptive conformal prediction methods. Apart from experiments on existing predictive inference benchmarks, we also demonstrate the state-of-the-art performance of the proposed methods on \textit{large-scale} tasks such as ImageNet classification and Cityscapes image segmentation.

****

Recon: Reducing Conflicting Gradients From the Root For Multi-Task Learning

Guangyuan SHI,Qimai Li,Wenlong Zhang,Jiaxin Chen,Xiao-Ming Wu

A fundamental challenge for multi-task learning is that different tasks may conflict with each other when they are solved jointly, and a cause of this phenomenon is conflicting gradients during optimization. Recent works attempt to mitigate the influence of conflicting gradients by directly altering the gradients based on some criteria. However, our empirical study shows that ``gradient surgery'' cannot effectively reduce the occurrence of conflicting gradients. In this paper, we take a different approach to reduce conflicting gradients from the root. In essence, we investigate the task gradients w.r.t. each shared network layer, select the layers with high conflict scores, and turn them to task-specific layers. Our experiments show that such a simple approach can greatly reduce the occurrence of conflicting gradients in the remaining shared layers and achieve better performance, with only a slight increase in model parameters in many cases. Our approach can be easily applied to improve various state-of-the-art methods including gradient manipulation methods and branched architecture search methods. Given a network architecture (e.g., ResNet18), it only needs to search for the conflict layers once, and the network can be modified to be used with different methods on the same or even different datasets to gain performance improvement. The source code is available at https://github.com/moukamisama/Recon.

****

OCD: Learning to Overfit with Conditional Diffusion Models

Shahar Lutati,Lior Wolf

We present a dynamic model in which the weights are conditioned on an input sample x and are learned to match those that would be obtained by finetuning a base model on x and its label y. This mapping between an input sample and network weights is shown to be approximated by a linear transformation of the sample distribution, which suggests that a denoising diffusion model can be suitable for this task. The diffusion model we therefore employ focuses on modifying a single layer of the base model and is conditioned on the input, activations, and output of this layer. Our experiments demonstrate the wide applicability of the method for image classification, 3D reconstruction, tabular data, and speech separation. Our code is attached as supplementary.

****

Measure the Predictive Heterogeneity

Jiashuo Liu,Jiayun Wu,Renjie Pi,Renzhe Xu,Xingxuan Zhang,Bo Li,Peng Cui

As an intrinsic and fundamental property of big data, data heterogeneity exists in a variety of real-world applications, such as in agriculture, sociology, health care, etc. For machine learning algorithms, the ignorance of data heterogeneity will significantly hurt the generalization performance and the algorithmic fairness, since the prediction mechanisms among different sub-populations are likely to differ. In this work, we focus on the data heterogeneity that affects the prediction of machine learning models, and first formalize the Predictive Heterogeneity, which takes into account the model capacity and computational constraints. We prove that it can be reliably estimated from finite data with PAC bounds even in high dimensions. Additionally, we propose the Information Maximization (IM) algorithm, a bi-level optimization algorithm, to explore the predictive heterogeneity of data. Empirically, the explored predictive heterogeneity provides insights for sub-population divisions in agriculture, sociology, and object recognition, and leveraging such heterogeneity benefits the out-of-distribution generalization performance.

****

On the robustness of self-supervised models for generative spoken language modeling

Itai Gat,Felix Kreuk,Ann Lee,Jade Copet,Gabriel Synnaeve,Emmanuel Dupoux,Yossi Adi

Self-supervised representations have been extensively studied for discriminative and generative tasks. However, their robustness capabilities have not been extensively investigated. This work focuses on self-supervised representations for spoken generative language models. First, we empirically demonstrate how current state-of-the-art speech representation models lack robustness to basic signal variations that do not alter the spoken information. To overcome this, we propose an effective and efficient method to learn robust self-supervised speech representation for generative spoken language modeling. The proposed approach is based on applying a set of signal transformations to the speech signal and optimizing the model using an iterative pseudo-labeling scheme. Our method significantly improves over the evaluated baselines when considering encoding metrics. We additi onally evaluate our method on the speech-to-speech translation task. We consider Spanish-English and French-English conversions and empirically demonstrate the benefits of following the proposed approach.

****

Non-equispaced Fourier Neural Solvers for PDEs

Haitao Lin,Yongjie Xu,Lirong Wu,Yufei Huang,Siyuan Li,Guojiang Zhao,Stan Z. Li

Solving partial differential equations is difficult. Recently proposed neural resolution-invariant models, despite their effectiveness and efficiency, usually require equispaced spatial points of data. However, sampling in spatial domain is sometimes inevitably non-equispaced in real-world systems, limiting their applicability. In this paper, we aim to propose a Non-equispaced Fourier PDE Solver (\textsc{NFS}) with adaptive interpolation on resampled equispaced points and a variant of Fourier Neural Operators as its components. Experimental results on complex PDEs demonstrate its advantages in accuracy and efficiency. Compared with the spatially-equispaced benchmark methods, it achieves superior performance with 42.85% improvements on MAE, and is able to handle non-equispaced data with a tiny loss of accuracy. Besides, to our best knowledge, \textsc{NFS} is the first ML-based method with mesh invariant inference ability to successfully model turbulent flows in non-equispaced scenarios, with a minor deviation of the error on unseen spatial points.

****

Time to augment self-supervised visual representation learning

Arthur Aubret,Markus R. Ernst,Céline Teulière,Jochen Triesch

Biological vision systems are unparalleled in their ability to learn visual representations without supervision. In machine learning, self-supervised learning (SSL) has led to major advances in forming object representations in an unsupervised fashion. Such systems learn representations invariant to augmentation operations over images, like cropping or flipping. In contrast, biological vision systems exploit the temporal structure of the visual experience during natural interactions with objects. This gives access to “augmentations” not commonly used in SSL, like watching the same object from multiple viewpoints or against different backgrounds. Here, we systematically investigate and compare the potential benefits of such time-based augmentations during natural interactions for learning object categories. Our results show that incorporating time-based augmentations achieves large performance gains over state-of-the-art image augmentations. Specifically, our analyses reveal that: 1) 3-D object manipulations drastically improve the learning of object categories; 2) viewing objects against changing backgrounds is important for learning to discard background-related information from the latent representation. Overall, we conclude that time-based augmentations during natural interactions with objects can substantially improve self-supervised learning, narrowing the gap between artificial and biological vision systems.

****

Probable Dataset Searching Method with Uncertain Dataset Information in Adjusting Architecture Hyper Parameter

Chen Yang,Jingyuan Wang

Different types of tasks with uncertain dataset information are studied because different parts of data may have different difficulties to achieve. For example, in unsupervised learning and domain adaptation, datasets are provided without label information because of the cost of human annotation. In deep learning, adjusting architecture hyper parameters is important for the model performance and is also time consuming, so we try to adjust hyper parameters in two types of uncertain dataset information:1, dataset labels are postponed to be obtained so hyper parameters need to be adjusted without complete dataset information. 2, hyper parameters are adjusted with a subset training dataset since training models with complete training dataset is time consuming. Here, we propose several loss functions to search for probable dataset when the complete dataset information is not obtained. The experiments on 9 real world data demonstrate the performance of our method.

****

Towards Lightweight, Model-Agnostic and Diversity-Aware Active Anomaly Detection
Xu Zhang,Yuan Zhao,Ziang Cui,Liqun Li,Shilin He,Qingwei Lin,Yingnong Dang,Saravan Rajmohan,Dongmei Zhang
Active Anomaly Discovery (AAD) is flourishing in the anomaly detection research area, which aims to incorporate analysts’ feedback into unsupervised anomaly detectors. However, existing AAD approaches usually prioritize the samples with the highest anomaly scores for user labeling, which hinders the exploration of anomalies that were initially ranked lower. Besides, most existing AAD approaches are specially tailored for a certain unsupervised detector, making it difficult to extend to other detection models. To tackle these problems, we propose a lightweight, model-agnostic and diversity-aware AAD method, named LMADA. In LMADA, we design a diversity-aware sample selector powered by Determinantal Point Process (DPP). It considers the diversity of samples in addition to their anomaly scores for feedback querying. Furthermore, we propose a model-agnostic tuner. It approximates diverse unsupervised detectors with a unified proxy model, based on which the feedback information is incorporated by a lightweight non-linear representation adjuster. Through extensive experiments on 8 public datasets, LMADA achieved 74% F1-Score improvement on average, outperforming other comparative AAD approaches. Besides, LMADA can also achieve significant performance boosting under any unsupervised detectors.
****
Switching One-Versus-the-Rest Loss to Increase Logit Margins for Adversarial Robustness
Sekitoshi Kanai,Shin'ya Yamaguchi,Masanori Yamada,Hiroshi Takahashi,Kentaro Ohno,Yasutoshi Ida
Adversarial training is a promising method to improve the robustness against adversarial attacks. To enhance its performance, recent methods impose high weights on the cross-entropy loss for important data points near the decision boundary. However, these importance-aware methods are vulnerable to sophisticated attacks, e.g., Auto-Attack. In this paper, we experimentally investigate the cause of their vulnerability via margins between logits for the true label and the other labels because they should be large enough to prevent the largest logit from being flipped by the attacks. Our experiments reveal that the histogram of the logit margins of naive adversarial training has two peaks. Thus, the levels of difficulty in increasing logit margins are roughly divided into two: difficult samples (small logit margins) and easy samples (large logit margins). On the other hand, only one peak near zero appears in the histogram of importance-aware methods, i.e., they reduce the logit margins of easy samples. To increase logit margins of difficult samples without reducing those of easy samples, we propose switching one-versus-the-rest loss (SOVR), which switches from cross-entropy to one-versus-the-rest loss (OVR) for difficult samples. We derive trajectories of logit margins for a simple problem and prove that OVR increases logit margins two times larger than the weighted cross-entropy loss. Thus, SOVR increases logit margins of difficult samples, unlike existing methods. We experimentally show that SOVR achieves better robustness against Auto-Attack than importance-aware methods.
****
Scaled Neural Multiplicative Model for Tractable Optimization
Drishya Giri,Kenneth Kuhn,Aravind Chiruvelli
Challenging decision problems in retail and beyond are often solved using the predict-then-optimize paradigm. An initial effort to develop and parameterize a model of an uncertain environment is followed by a separate effort to identify the best possible solution of an optimization problem. Linear models are often used to ensure optimization problems are tractable. Remarkably accurate Deep Neural Network (DNN) models have recently been developed for various prediction tasks. Such models have been shown to scale to large datasets without loss of accuracy and with good computational performance. It can, however, be challenging to formulate tractable optimization problems based on DNN models. In this work we consider the problem of shelf space allocation for retail stores using DNN models. We highlight the trade-off between predictive performance and the tractability of optimization problems. We introduce a Scaled Neural Multiplicative Model (SNMM)

with shape constraints for demand learning that leads to a tractable optimization formulation. Although, this work focuses on a specific application, the formulation of the models are general enough such that they can be extended to many real world applications.
****

Quasi-Taylor Samplers for Diffusion Generative Models based on Ideal Derivatives
Hideyuki Tachibana,Mocho Go,Muneyoshi Inahara,Yotaro Katayama,Yotaro Watanabe
Diffusion generative models have emerged as a new challenger to popular deep neural generative models such as GANs, but have the drawback that they often require a huge number of neural function evaluations (NFEs) during synthesis unless some sophisticated sampling strategies are employed. This paper proposes new efficient samplers based on the numerical schemes derived by the familiar Taylor expansion, which directly solves the ODE/SDE of interest. In general, it is not easy to compute the derivatives that are required in higher-order Taylor schemes, but in the case of diffusion models, this difficulty is alleviated by the trick that the authors call ``ideal derivative substitution,'' in which the higher-order derivatives are replaced by tractable ones. To derive ideal derivatives, the authors argue the ``single point approximation,'' in which the true score function is approximated by a conditional one, holds in many cases, and considered the derivatives of this approximation. Applying thus obtained new quasi-Taylor samplers to image generation tasks, the authors experimentally confirmed that the proposed samplers could synthesize plausible images in small number of NFEs, and that the performance was better or at the same level as DDIM and Runge-Kutta methods. The paper also argues the relevance of the proposed samplers to the existing ones mentioned above.
****

Group-oriented Cooperation in Multi-Agent Reinforcement Learning
Yifan Zang,Jinmin He,Kai Li,Haobo Fu,QIANG FU,Junliang Xing,Jian Cheng
Grouping is ubiquitous in natural systems and is essential for promoting efficiency in team coordination. This paper introduces the concept of grouping into multi-agent reinforcement learning (MARL) and provides a novel formulation of Group-oriented MARL (GoMARL). In contrast to existing approaches that attempt to directly learn the complex relationship between the joint action-values and individual values, we empower groups as a bridge to model the connection between a small set of agents and encourage cooperation among them, thereby improving the efficiency of the whole team. In particular, we factorize the joint action-values as a combination of group-wise values, which guide agents to improve their policies in a fine-grained fashion. We propose a flexible grouping mechanism inspired by variable selection and sparse regularization to generate dynamic groups and group action-values. We further propose a hierarchical control for policy learning that drives the agents in the same group to specialize in similar policies and possess diversified strategies for various groups. Extensive experiments on a challenging set of StarCraft II micromanagement tasks and Google Research Football scenarios verify our method's effectiveness and learning efficiency. Detailed component studies show how grouping works and enhances performance.
****

Exploring Temporally Dynamic Data Augmentation for Video Recognition
Taeoh Kim,Jinhyung Kim,Minho Shim,Sangdoo Yun,Myunggu Kang,Dongyoon Wee,Sangyoun Lee
Data augmentation has recently emerged as an essential component of modern training recipes for visual recognition tasks.
However, data augmentation for video recognition has been rarely explored despite its effectiveness.
Few existing augmentation recipes for video recognition naively extend the image augmentation methods by applying the same operations to the whole video frames.
Our main idea is that the magnitude of augmentation operations for each frame needs to be changed over time to capture the real-world video's temporal variations.
These variations should be generated as diverse as possible using fewer additional hyper-parameters during training.

Through this motivation, we propose a simple yet effective video data augmentation framework, DynaAugment.

The magnitude of augmentation operations on each frame is changed by an effective mechanism, Fourier Sampling that parameterizes diverse, smooth, and realistic temporal variations.

DynaAugment also includes an extended search space suitable for video for automatic data augmentation methods.

DynaAugment experimentally demonstrates that there are additional performance rooms to be improved from static augmentations on diverse video models.

Specifically, we show the effectiveness of DynaAugment on various video datasets and tasks: large-scale video recognition (Kinetics-400 and Something-Something-v2), small-scale video recognition (UCF-101 and HMDB-51), fine-grained video recognition (Diving-48 and FineGym), video action segmentation on Breakfast, video action localization on THUMOS'14, and video object detection on MOT17Det.

****

Agent Prioritization with Interpretable Relation for Trajectory Prediction

Manh Huynh,Hengbo Ma,Gita Alaghband,Chiho Choi

In this paper, we present a novel multi-agent trajectory prediction model, which discovers interpretable relations among agents and prioritize agent's motion. Different from existing approaches, our interpretable design is inspired by the fundamental navigation and motion functions of agent movements, which represent 'where' and 'how' the agents move in the scenes. Specifically, it generates the relation matrix, where each element indicates the motion impact from one to another. In addition, in highly interactive scenarios, one agent may implicitly gain higher priority to move, while the motion of other agents may be impacted by the prioritized agents with higher priority (e.g., a vehicle stopping or reducing its speed due to crossing pedestrians). Based on this intuition, we design a novel motion prioritization module to learn the agent motion priorities based on the inferred relation matrix. Then, a decoder is proposed to sequentially predict and iteratively update the future trajectories of each agent based on their priority orders and the learned relation structures. We first demonstrate the effectiveness of our prediction model on simulated Charged Particles dataset. Next, extensive evaluations are performed on commonly-used datasets for robot navigation, human-robot interactions, and autonomous agents: real-world NBA basketball and INTERACTION. Finally, we show that the proposed model outperforms other state-of-the-art relation based methods, and is capable to infer interpretable, meaningful relations among agents.

****

z-SignFedAvg: A Unified Stochastic Sign-based Compression for Federated Learning

Zhiwei Tang,Yanmeng Wang,Tsung-Hui Chang

Federated Learning (FL) is a promising privacy-preserving distributed learning paradigm but suffers from high communication cost when training large-scale machine learning models. Sign-based methods, such as SignSGD \citep{bernstein2018signsgd}, have been proposed as a biased gradient compression technique for reducing the communication cost. However, sign-based algorithms could diverge under heterogeneous data, which thus motivated the development of advanced techniques, such as the error-feedback method and stochastic sign-based compression, to fix this issue.

Nevertheless, these methods still suffer from slower convergence rates. Besides, none of them allows multiple local SGD updates like FedAvg \citep{mcmahan2017communication}. In this paper, we propose a novel noisy perturbation scheme with a general symmetric noise distribution for sign-based compression, which not only allows one to flexibly control the tradeoff between gradient bias and convergence performance, but also provides a unified viewpoint to existing stochastic sign-based methods. More importantly, we propose the very first sign-based FedAvg algorithm (z-SignFedAvg). Theoretically, we show that z-SignFedAvg achieves a faster convergence rate than existing sign-based methods and, under the uniformly distributed noise, can enjoy the same convergence rate as its uncompressed counterpart. Last but not the least, we remark that adding random noise to the local gradients has a double benefit: it protects the clients' privacy by, e.g., the Differential Privacy. Extensive experiments are conducted to demonstrate that the z-SignFedAvg can achieve competitive empirical performance on real datasets.

****

Transfer Learning with Pre-trained Conditional Generative Models
Shin'ya Yamaguchi,Sekitoshi Kanai,Atsutoshi Kumagai,Daiki Chijiwa,Hisashi Kashima

Transfer learning is crucial in training deep neural networks on new target tasks. Current transfer learning methods always assume at least one of (i) source and target task label spaces overlap, (ii) source datasets are available, and (iii) target network architectures are consistent with source ones. However, holding these assumptions is difficult in practical settings because the target task rarely has the same labels as the source task, the source dataset access is restricted due to storage costs and privacy, and the target architecture is often specialized to each task. To transfer source knowledge without these assumptions, we propose a transfer learning method that uses deep generative models and is composed of the following two stages: pseudo pre-training (PP) and pseudo semi-supervised learning (P-SSL). PP trains a target architecture with an artificial dataset synthesized by using conditional source generative models. P-SSL applies SSL algorithms to labeled target data and unlabeled pseudo samples, which are generated by cascading the source classifier and generative models to condition them with target samples. Our experimental results indicate that our method can outperform the baselines of scratch training and knowledge distillation.

****

DECN: Evolution Inspired Deep Convolution Network for Black-box Optimization
Liu Penghui,Kai Wu,Jing Liu

We design a deep evolutionary convolution network (DECN) for continuous black-box optimization to force the random population to move near the optimal solution, which is the goal of population-based optimization, such as evolutionary algorithm and evolutionary strategy. DECN is composed of two modules: convolution-based reasoning module (CRM) and selection module (SM), to move from hand-designed searching strategies to learned searching strategies in population-based optimization. CRM produces a population that is closer to the optimal solution based on the convolution operators, and SM removes poor solutions. We also design a proper loss function to support the training of DECN. The experimental results on unconstrained continuous optimization problems show that DECN can learn searching strategies and surpass population-based baselines. Moreover, DECN obtains good performance when transferred to optimization problems unseen during the training stage. In addition, DECN is friendly to the acceleration with Graphics Processing Units (GPUs) and runs 102 times faster than unaccelerated EA when evolving 32 populations, each containing 6400 individuals.

****

Q-Pensieve: Boosting Sample Efficiency of Multi-Objective RL Through Memory Sharing of Q-Snapshots
Wei Hung,Bo Kai Huang,Ping-Chun Hsieh,Xi Liu

Many real-world continuous control problems are in the dilemma of weighing the pros and cons, multi-objective reinforcement learning (MORL) serves as a generic framework of learning control policies for different preferences over objectives. However, the existing MORL methods either rely on multiple passes of explicit search for finding the Pareto front and therefore are not sample-efficient, or utilizes a shared policy network for coarse knowledge sharing among policies. To boost the sample efficiency of MORL, we propose Q-Pensieve, a policy improvement scheme that stores a collection of Q-snapshots to jointly determine the policy update direction and thereby enables data sharing at the policy level. We show that Q-Pensieve can be naturally integrated with soft policy iteration with convergence guarantee. To substantiate this concept, we propose the technique of Q replay buffer, which stores the learned Q-networks from the past iterations, and arrive at a practical actor-critic implementation. Through extensive experiments and an ablation study, we demonstrate that with much fewer samples, th e proposed algorithm can outperform the benchmark MORL methods on a variety of MORL benchmark tasks.

****

Optformer: Beyond Transformer for Black-box Optimization

Xiaobin Li,Kai Wu,Xiaoyu Zhang,Handing Wang,Jing Liu

We design a novel Transformer for continuous unconstrained black-box optimization, called Optformer. Inspired by the similarity between Vision Transformer and evolutionary algorithms (EAs), we modify Tansformer's multi-head self-attention layer, feed-forward network, and residual connection to implement the functions of crossover, mutation, and selection operators. Moreover, we devise an iterated mode to generate and survive potential solutions like EAs. Optformer establishes a mapping from the random population to the optimal population. Compared to baselines, such as EAs, Bayesian optimization, and the learning-to-optimize method, Optformer shows the top performance in six black-box functions and one real-world application. We also find that untrained Optformer can also achieve good performance.

****

Deformable Graph Transformer

Jinyoung Park,Seongjun Yun,Hyeonjin Park,Jaewoo Kang,Jisu Jeong,Kyung-Min Kim,Jung-Woo Ha,Hyunwoo J. Kim

Transformer-based models have recently shown success in representation learning on graph-structured data beyond natural language processing and computer vision. However, the success is limited to small-scale graphs due to the drawbacks of full dot-product attention on graphs such as the quadratic complexity with respect to the number of nodes and message aggregation from enormous irrelevant nodes. To address these issues, we propose Deformable Graph Transformer (DGT) that performs sparse attention via dynamically sampled relevant nodes for efficiently handling large-scale graphs with a linear complexity in the number of nodes. Specifically, our framework first constructs multiple node sequences with various criteria to consider both structural and semantic proximity. Then, combining with our learnable Katz Positional Encodings, the sparse attention is applied to the node sequences for learning node representations with a significantly reduced computational cost. Extensive experiments demonstrate that our DGT achieves state-of-the-art performance on 7 graph benchmark datasets with 2.5 ~ 449 times less computational cost compared to transformer-based graph models with full attention.

****

Exact manifold Gaussian Variational Bayes

Martin Magris,Mostafa Shabani,Alexandros Iosifidis

We propose an optimization algorithm for Variational Inference (VI) in complex models. Our approach relies on natural gradient updates where the variational space is a Riemann manifold. We develop an efficient algorithm for Gaussian Variational Inference that implicitly satisfies the positive definite constraint on the variational covariance matrix. Our Exact manifold Gaussian Variational Bayes (EMGVB) provides exact but simple update rules and is straightforward to implement. Due to its black-box nature, EMGVB stands as a ready-to-use solution for VI in complex models. Over five datasets, we empirically validate our feasible approach on different statistical, econometric, and deep learning models, discussing its performance with respect to baseline methods.

****

Variance-Aware Sparse Linear Bandits

Yan Dai,Ruosong Wang,Simon Shaolei Du

It is well-known that for sparse linear bandits, when ignoring the dependency on sparsity which is much smaller than the ambient dimension, the worst-case minimax regret is $\widetilde{\Theta}\left(\sqrt{dT}\right)$ where d is the ambient dimension and T is the number of rounds. On the other hand, in the benign setting where there is no noise and the action set is the unit sphere, one can use divide-and-conquer to achieve $\widetilde{\mathcal O}(1)$ regret, which is (nearly) independent of d and T. In this paper, we present the first variance-aware regret guarantee for sparse linear bandits: $\widetilde{\mathcal O}\left(\sqrt{d\sum_{t=1}^T \sigma_t^2} + 1\right)$, where σ_t^2 is the variance of the noise at the t-th round. This bound naturally interpolates the regret bounds for the worst-case constant-variance regime (i.e., $\sigma_t \equiv \Omega(1)$) and the benign deterministic regimes (i.e., $\sigma_t \equiv 0$). To achieve this variance-aware regret guarantee, we develop a general framework that converts any variance-aware linear bandit algorithm to a variance-aware algorithm for sparse linear bandits in a "black-box" manner. Specifically, we take two recent algorithms as black boxes to illustrate that the claimed bounds indeed hold, where the first algorithm can handle unknown-variance cases and the second one is more efficient.

****

Multi-Treatment Effect Estimation with Proxy: Contrastive Learning and Rank Weighting

Minqin Zhu,Anpeng Wu,Ruoxuan Xiong,Kun Kuang

We study the treatment effect estimation problem for continuous and multi-dimensional treatments, in the setting with unobserved confounders, but high-dimension proxy variables for unobserved confounders are available. Existing methods either directly adjust the relationship between observed covariates and treatments or recover the hidden confounders by probabilistic models. However, they either rely on a correctly specified treatment assignment model or require strong prior of the unobserved confounder distribution. To relax these requirements, we propose a Contrastive Regularizer (CR) to learn the proxy representation that contains all the relevant information in unobserved confounders. Based on the CR, we propose a novel ranked weighting method (Rw) to de-bias the treatment assignment. Combining Cr and Rw, we propose a neural network framework named CRNet to estimate the effects of multiple continuous treatments under unobserved confounders, evaluated by the Average Dose-Response Function. Empirically, we demonstrate that CRNet achieves state-of-the-art performance on both synthetic and semi-synthetic datasets.

****

CircNet: Meshing 3D Point Clouds with Circumcenter Detection

Huan Lei,Ruitao Leng,Liang Zheng,Hongdong Li

Reconstructing 3D point clouds into triangle meshes is a key problem in computational geometry and surface reconstruction. Point cloud triangulation solves this problem by providing edge information to the input points. Since no vertex interpolation is involved, it is beneficial to preserve sharp details on the surface. Taking advantage of learning-based techniques in triangulation, existing methods enumerate the complete combinations of candidate triangles, which is both complex and inefficient. In this paper, we leverage the duality between a triangle and its circumcenter, and introduce a deep neural network that detects the circumcenters to achieve point cloud triangulation. Specifically, we introduce multiple anchor priors to divide the neighborhood space of each point. The neural network then learns to predict the presences and locations of circumcenters under the guidance of those anchors. We extract the triangles dual to the detected circumcenters to form a primitive mesh, from which an edge-manifold mesh is produced via simple post-processing. Unlike existing learning-based triangulation methods, the proposed method bypasses an exhaustive enumeration of triangle combinations and local surface parameterization. We validate the efficiency, generalization, and robustness of our method on prominent datasets of both watertight and open surfaces. The code and trained models are provided at \url{https://github.com/Ruitao-L/CircNet}.

****

In-sample Actor Critic for Offline Reinforcement Learning

Hongchang Zhang,Yixiu Mao,Boyuan Wang,Shuncheng He,Yi Xu,Xiangyang Ji

Offline reinforcement learning suffers from out-of-distribution issue and extrapolation error. Most methods penalize the out-of-distribution state-action pairs or regularize the trained policy towards the behavior policy but cannot guarantee to get rid of extrapolation error. We propose In-sample Actor Critic (IAC) which utilizes sampling-importance resampling to execute in-sample policy evaluation. IAC only uses the target Q-values of the actions in the dataset to evaluate the trained policy, thus avoiding extrapolation error. The proposed method per forms unbiased policy evaluation and has a lower variance than importance sampling in many cases. Empirical results show that IAC obtains competitive performance compared to the state-of-the-art methods on Gym-MuJoCo locomotion domains and much more challenging AntMaze domains.

****

Leveraging Future Relationship Reasoning for Vehicle Trajectory Prediction

Daehee Park,Hobin Ryu,Yunseo Yang,Jegyeong Cho,Jiwon Kim,Kuk-Jin Yoon

Understanding the interaction between multiple agents is crucial for realistic vehicle trajectory prediction.

Existing methods have attempted to infer the interaction from the observed past trajectories of agents using pooling, attention, or graph-based methods, which rely on a deterministic approach.

However, these methods can fail under complex road structures, as they cannot predict various interactions that may occur in the future.

In this paper, we propose a novel approach that uses lane information to predict a stochastic future relationship among agents.

To obtain a coarse future motion of agents, our method first predicts the probability of lane-level waypoint occupancy of vehicles.

We then utilize the temporal probability of passing adjacent lanes for each agent pair, assuming that agents passing adjacent lanes will highly interact.

We also model the interaction using a probabilistic distribution, which allows for multiple possible future interactions.

The distribution is learned from the posterior distribution of interaction obtained from ground truth future trajectories.

We validate our method on popular trajectory prediction datasets: nuScenes and Argoverse.

The results show that the proposed method brings remarkable performance gain in prediction accuracy, and achieves state-of-the-art performance in long-term prediction benchmark dataset.

****

DeepTime: Deep Time-index Meta-learning for Non-stationary Time-series Forecasting

Gerald Woo,Chenghao Liu,Doyen Sahoo,Akshat Kumar,Steven Hoi

Advances in I.T. infrastructure has led to the collection of longer sequences of time-series. Such sequences are typically non-stationary, exhibiting distribution shifts over time -- a challenging scenario for the forecasting task, due to the problems of covariate shift, and conditional distribution shift. In this paper, we show that deep time-index models possess strong synergies with a meta-learning formulation of forecasting, displaying significant advantages over existing neural forecasting methods in tackling the problems arising from non-stationarity. These advantages include having a stronger smoothness prior, avoiding the problem of covariate shift, and having better sample efficiency. To this end, we propose DeepTime, a deep time-index model trained via meta-learning. Extensive experiments on real-world datasets in the long sequence time-series forecasting setting demonstrate that our approach achieves competitive results with state-of-the-art methods, and is highly efficient. Code is attached as supplementary material, and will be publicly released.

****

Non-Parametric State-Space Models: Identifiability, Estimation and Forecasting

Chenghao Liu,Weiran Yao,Steven Hoi,Kun Zhang

State-space models (SSMs) provide a standard methodology for time series analysis and prediction. While recent works utilize nonlinear functions to parameterize the transition and emission processes to enhance their expressivity, the form of additive noise still limits their applicability in real-world scenarios. In this work, we propose a general formulation of SSMs with a completely non-parametric transition model and a flexible emission model which can account for sensor distortion. Besides, to deal with more general scenarios (e.g., non-stationary time series), we add a higher level model to capture time-varying characteristics of the process.

Interestingly, we find that even though the proposed model is remarkably flexibl e, the latent processes are generally identifiable. Given this, we further propose the corresponding estimation procedure and make use of it for the forecasting task. Our model can recover the latent processes and their relations from observed sequential data. Accordingly, the proposed procedure can also be viewed as a method for causal representation learning. We argue that forecasting can benefit from causal representation learning, since the estimated latent variables are generally identifiable. Empirical comparisons on various datasets validate that our model could not only reliably identify the latent processes from the observed data, but also consistently outperform baselines in the forecasting task.

****

ETSformer: Exponential Smoothing Transformers for Time-series Forecasting

Gerald Woo,Chenghao Liu,Doyen Sahoo,Akshat Kumar,Steven Hoi

Transformers have recently been actively studied for time-series forecasting. While often showing promising results in various scenarios, traditional Transformers are not designed to fully exploit the characteristics of time-series data and thus suffer some fundamental limitations, e.g., they are generally not decomposable or interpretable, and are neither effective nor efficient for long-term forecasting. In this paper, we propose ETSformer, a novel time-series Transformer architecture, which exploits the principle of exponential smoothing methods in improving Transformers for time-series forecasting. Specifically, ETSformer leverages a novel level-growth-seasonality decomposed Transformer architecture which leads to more interpretable and disentangled decomposed forecasts. We further propose two novel attention mechanisms -- the exponential smoothing attention and frequency attention, which are specially designed to overcome the limitations of the vanilla attention mechanism for time-series data. Extensive experiments on various time-series benchmarks validate the efficacy and advantages of the proposed method. Code is attached in the supplementary material, and will be made publicly available.

****

LMSeg: Language-guided Multi-dataset Segmentation

Qiang Zhou,Yuang Liu,Chaohui Yu,Jingliang Li,Zhibin Wang,Fan Wang

It’s a meaningful and attractive topic to build a general and inclusive segmentation model that can recognize more categories in various scenarios. A straightforward way is to combine the existing fragmented segmentation datasets and train a multi-dataset network. However, there are two major issues with multi-dataset segmentation: (i) the inconsistent taxonomy demands manual reconciliation to construct a unified taxonomy; (ii) the inflexible one-hot common taxonomy causes time-consuming model retraining and defective supervision of unlabeled categories. In this paper, we investigate the multi-dataset segmentation and propose a scalable Language-guided Multi-dataset Segmentation framework, dubbed LMSeg, which supports both semantic and panoptic segmentation. Specifically, we introduce a pretrained text encoder to map the category names to a text embedding space as a unified taxonomy, instead of using inflexible one-hot label. The model dynamically aligns the segment queries with the category embeddings. Instead of relabeling each dataset with the unified taxonomy, a category-guided decoding module is designed to dynamically guide predictions to each dataset’s taxonomy. Furthermore, we adopt a dataset-aware augmentation strategy that assigns each dataset a specific image augmentation pipeline, which can suit the properties of images from different datasets. Extensive experiments demonstrate that our method achieves significant improvements on four segmentation datasets and three panoptic datasets, while the ablation study evaluates the effectiveness of each component.

****

Horizon-Free Reinforcement Learning for Latent Markov Decision Processes

Runlong Zhou,Ruosong Wang,Simon Shaolei Du

We study regret minimization for reinforcement learning (RL) in Latent Markov Decision Processes (LMDPs) with context in hindsight. We design a novel model-based algorithmic framework which can be instantiated with both a model-optimistic and a value-optimistic solver. We prove an $\widetilde{O}\left(\sqrt{M \Gamma S A

K}\right)$ regret bound where M is the number of contexts, S is the number of states, A is the number of actions, K is the number of episodes, and $\Gamma \le S$ is the maximum transition degree of any state-action pair. The regret bound only scales logarithmically with the planning horizon, thus yielding the first (nearly) horizon-free regret bound for LMDP. Key in our proof is an analysis of the total variance of alpha vectors, which is carefully bounded by a recursion-based technique. We complement our positive result with a novel $\Omega\left(\sqrt{M S A K}\right)$ regret lower bound with $\Gamma = 2$, which shows our upper bound minimax optimal when Γ is a constant. Our lower bound relies on new constructions of hard instances and an argument based on the symmetrization technique from theoretical computer science, both of which are technically different from existing lower bound proof for MDPs, and thus can be of independent interest.

****

Learning Invariant Features for Online Continual Learning

Yiduo Guo,Bing Liu,Dongyan Zhao

It has been shown recently that learning only discriminative features that are sufficient to separate the classes in a task using a traditional learning method has a major shortcoming for continual learning (CL). This is because many features that are not learned may be necessary for distinguishing classes of some future tasks. When such a future task arrives, these features have to be learned by updating the network, which causes catastrophic forgetting (CF). A recent work on online CL showed that if the learning method can learn as many features as possible from each class, called holistic representations, CF can be significantly reduced to achieve a large performance gain. This paper argues that learning only holistic representations is still insufficient. The learned representations should also be invariant and those features that are present in the data but are irrelevant to the class (e.g., the background information) should be ignored for better generalization across tasks. This new condition further boosts the performance significantly. This paper proposes several strategies and a loss to learn holistic and invariant representations and evaluates their effectiveness in online CL.

****

RoPAWS: Robust Semi-supervised Representation Learning from Uncurated Data

Sangwoo Mo,Jong-Chyi Su,Chih-Yao Ma,Mido Assran,Ishan Misra,Licheng Yu,Sean Bell

Semi-supervised learning aims to train a model using limited labels. State-of-the-art semi-supervised methods for image classification such as PAWS rely on self-supervised representations learned with large-scale unlabeled but curated data. However, PAWS is often less effective when using real-world unlabeled data that is uncurated, e.g., contains out-of-class data. We propose RoPAWS, a robust extension of PAWS that can work with real-world unlabeled data. We first reinterpret PAWS as a generative classifier that models densities using kernel density estimation. From this probabilistic perspective, we calibrate its prediction based on the densities of labeled and unlabeled data, which leads to a simple closed-form solution from the Bayes' rule. We demonstrate that RoPAWS significantly improves PAWS for uncurated Semi-iNat by +5.3% and curated ImageNet by +0.4%.

****

Treeformer: Dense Gradient Trees for Efficient Attention Computation

Lovish Madaan,Srinadh Bhojanapalli,Himanshu Jain,Prateek Jain

Standard inference and training with transformer based architectures scale quadratically with input sequence length. This is prohibitively large for a variety of applications especially in web-page translation, query-answering etc. Consequently, several approaches have been developed recently to speedup attention computation by enforcing different attention structures such as sparsity, low-rank, approximating attention using kernels. In this work, we view attention computation as that of nearest neighbor retrieval, and use decision tree based hierarchical navigation to reduce the retrieval cost per query token from linear in sequence length to nearly logarithmic. Based on such hierarchical navigation, we design Treeformer which can use one of two efficient attention layers -- TF-Attention and TC-Attention. TF-Attention computes the attention in a fine-grained style, w hile TC-Attention is a coarse attention layer which also ensures that the gradients are "dense". To optimize such challenging discrete layers, we propose a two-level bootstrapped training method. Using extensive experiments on standard NLP benchmarks, especially for long-sequences, we demonstrate that our Treeformer architecture can be almost as accurate as baseline Transformer while using 30x lesser FLOPs in the attention layer. Compared to Linformer, the accuracy can be as much as 12% higher while using similar FLOPs in the attention layer.

****

ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection

Chenyang ZHAO,Antoni B. Chan

We propose the Gradient-weighted Object Detector Activation Mapping (Grad-ODAM), a visualized explanation technique for interpreting the predictions of object detectors. Utilizing the gradients of detector targets flowing into the intermediate feature maps, Grad-ODAM produces heat maps that show the influence of regions on the detector's decision. Compared to previous classification activation mapping works, Grad-ODAM generates instance-specific explanations rather than class-specific ones. We show that Grad-ODAM is applicable to both one-stage detectors such as FCOS and two-stage detectors such as Faster R-CNN, and produces higher-quality visual explanations than the state-of-the-art both effectively and efficiently. We next propose a training scheme, ODAM-Train, to improve the explanation ability on object discrimination of the detector through encouraging consistency between explanations for detections on the same object, and distinct explanations for detections on different objects. Based on the heat maps produced by Grad-ODAM with ODAM-Train, we propose ODAM-NMS, which considers the information of the model's explanation for each prediction to distinguish the duplicate detected objects. We present a detailed analysis of the visualized explanations of detectors and carry out extensive experiments to validate the effectiveness of the proposed ODAM.

****

Understanding Curriculum Learning in Policy Optimization for Online Combinatorial Optimization

Runlong Zhou,Yuandong Tian,Yi Wu,Simon Shaolei Du

Over the recent years, reinforcement learning (RL) starts to show promising results in tackling combinatorial optimization (CO) problems, in particular when coupled with curriculum learning to facilitate training. Despite emerging empirical evidence, theoretical study on why RL helps is still at its early stage. This paper presents the first systematic study on policy optimization methods for online CO problems. We show that online CO problems can be naturally formulated as latent Markov Decision Processes (LMDPs), and prove convergence bounds on natural policy gradient (NPG) for solving LMDPs. Furthermore, our theory explains the benefit of curriculum learning: it can find a strong sampling policy and reduce the distribution shift, a critical quantity that governs the convergence rate in our theorem. For a canonical online CO problem, Secretary Problem, we formally prove that distribution shift is reduced exponentially with curriculum learning even if the curriculum is randomly generated. Our theory also shows we can simplify the curriculum learning scheme used in prior work from multi-step to single-step. Lastly, we provide extensive experiments on Secretary Problem and Online Knapsack to verify our findings.

****

Toward Adversarial Training on Contextualized Language Representation

Hongqiu Wu,Yongxiang Liu,Hanwen Shi,hai zhao,Min Zhang

Beyond the success story of adversarial training (AT) in the recent text domain on top of pre-trained language models (PLMs), our empirical study showcases the inconsistent gains from AT on some tasks, e.g. commonsense reasoning, named entity recognition. This paper investigates AT from the perspective of the contextualized language representation outputted by PLM encoders. We find the current AT attacks lean to generate sub-optimal adversarial examples that can fool the decoder part but have a minor effect on the encoder. However, we find it necessary to effectively deviate the latter one to allow AT to gain. Based on the observati on, we propose simple yet effective \textit{Contextualized representation-Adversarial Training} (CreAT), in which the attack is explicitly optimized to deviate the contextualized representation of the encoder. It allows a global optimization of adversarial examples that can fool the entire model. We also find CreAT gives rise to a better direction to optimize the adversarial examples, to let them less sensitive to hyperparameters. Compared to AT, CreAT produces consistent performance gains on a wider range of tasks and is proven to be more effective for language pre-training where only the encoder part is kept for downstream tasks. We achieve the new state-of-the-art performances on a series of challenging benchmarks, e.g. AdvGLUE (59.1 $ \rightarrow $ 61.1), HellaSWAG (93.0 $ \rightarrow $ 94.9), ANLI (68.1 $ \rightarrow $ 69.3).

****

Efficient Method for Bi-level Optimization with Non-smooth Lower-Level Problem

Wanli Shi,Bin Gu

Bi-level optimization plays a key role in a lot of machine learning applications. Existing state-of-the-art bi-level optimization methods are limited to smooth or some specific non-smooth lower-level problems. Therefore, achieving an efficient algorithm for the bi-level problems with a generalized non-smooth lower-level objective is still an open problem. To address this problem, in this paper, we propose a new bi-level optimization algorithm based on smoothing and penalty techniques. Using the theory of generalized directional derivative, we derive new conditions for the bilevel optimization problem with nonsmooth, perhaps non-Lipschitz lower-level problem, and prove our method can converge to the points satisfying these conditions. We also compare our method with existing state-of-the-art bi-level optimization methods and demonstrate that our method is superior to the others in terms of accuracy and efficiency.

****

Estimating Riemannian Metric with Noise-Contaminated Intrinsic Distance

Jiaming Qiu,Xiongtao Dai

We extend metric learning by studying the Riemannian manifold structure of the underlying data space induced by dissimilarity measures between data points. The key quantity of interest here is the Riemannian metric, which characterizes the Riemannian geometry and defines straight lines and derivatives on the manifold. Being able to estimate the Riemannian metric allows us to gain insights into the underlying manifold and compute geometric features such as the geodesic curves. We model the observed dissimilarity measures as noisy responses generated from a function of the intrinsic geodesic distance between data points. A new local regression approach is proposed to learn the Riemannian metric tensor and its derivatives based on a Taylor expansion for the squared geodesic distances. Our framework is general and accommodates different types of responses, whether they are continuous, binary, or comparative, extending the existing works which consider a single type of response at a time. We develop theoretical foundation for our method by deriving the rates of convergence for the asymptotic bias and variance of the estimated metric tensor. The proposed method is shown to be versatile in simulation studies and a real data application involving taxi trip time in New York City.

****

In Search of Smooth Minima for Purifying Backdoor in Deep Neural Networks

Nazmul Karim,Abdullah Al Arafat,Umar Khalid,Zhishan Guo,Nazanin Rahnavard

The success of a deep neural network (DNN) heavily relies on the details of the training scheme; e.g., training data, architectures, hyper-parameters, etc. Recent backdoor attacks suggest that an adversary can take advantage of such training details and compromise the integrity of DNN. Our studies show that a backdoor model is usually optimized to a bad local minima, i.e., sharper minima as compared to a benign model. Intuitively, backdoor can be purified by re-optimizing the model to a smoother minima through fine-tuning with a few clean validation data. However, fine-tuning all DNN parameters often requires huge computational costs as well as sub-par clean test performance. To address this concern, we propose a novel backdoor purification technique—N atural G radient Fine-tuning (NGF)—which focuses on removing backdoor by fine-tuning only one layer. Specifically, NGF utilizes a loss surface geometry-aware optimizer that can successfully overcome the challenge of reaching a smooth minima under one-layer optimization scenario. To enhance the generalization performance of our proposed method, we introduce a clean data distribution-aware regularizer based on the knowledge of loss surface curvature matrix, i.e., Fisher Information Matrix. To validate the effectiveness of our method, we conduct extensive experimentation with four different datasets—CIFAR10, GTSRB, Tiny-ImageNet, and ImageNet; as well as 11 recent backdoor attacks, e.g., Blend, Dynamic, Clean Label, etc. NGF achieves state-of-the-art performance in most of these benchmarks.

****

Joint Gaussian Mixture Model for Versatile Deep Visual Model Explanation

Zhouyang Xie,Duanbing Chen

Post-hoc explanations of deep neural networks improve human understanding on the learned representations, decision-making process and uncertainty of the model with faithfulness. Explaining deep convolutional neural networks(DCNN) is especially challenging, due to the high dimensionality of deep features and the complexity of model inference. Most post-hoc explaining methods serve a single form of explanation, restricting the diversity and consistency of the explanation. This paper proposes joint Gaussian mixture model(JGMM), a probabilistic model jointly models inter-layer deep features and produces faithful and consistent post-hoc explanations. JGMM explains deep features by Gaussian mixture model and inter-layer deep feature relations by posterior distribution on the latent component variables. JGMM enables a versatile explaining framework that unifies interpretable proxy model, global or local explanatory example generation or mining. Experiments are performed on various DCNN image classifiers in comparison with other explaining methods. It shows that JGMM can efficiently produce versatile, consistent, faithful and understandable explanations.

****

Gromov-Wasserstein Autoencoders

Nao Nakagawa,Ren Togo,Takahiro Ogawa,Miki Haseyama

Variational Autoencoder (VAE)-based generative models offer flexible representation learning by incorporating meta-priors, general premises considered beneficial for downstream tasks. However, the incorporated meta-priors often involve ad-hoc model deviations from the original likelihood architecture, causing undesirable changes in their training. In this paper, we propose a novel representation learning method, Gromov-Wasserstein Autoencoders (GWAE), which directly matches the latent and data distributions using the variational autoencoding scheme. Instead of likelihood-based objectives, GWAE models minimize the Gromov-Wasserstein (GW) metric between the trainable prior and given data distributions. The GW metric measures the distance structure-oriented discrepancy between distributions even with different dimensionalities, which provides a direct measure between the latent and data spaces. By restricting the prior family, we can introduce meta-priors into the latent space without changing their objective. The empirical comparisons with VAE-based models show that GWAE models work in two prominent meta-priors, disentanglement and clustering, with their GW objective unchanged.

****

Localized Graph Contrastive Learning

Hengrui Zhang,Qitian Wu,Yu Wang,Shaofeng Zhang,Junchi Yan,Philip S. Yu

Contrastive learning methods based on InfoNCE loss are popular in node representation learning tasks on graph-structured data. However, its reliance on data augmentation and its quadratic computational complexity might lead to inconsistency and inefficiency problems. To mitigate these limitations, in this paper, we introduce a simple yet effective contrastive model named Localized Graph Contrastive Learning (Local-GCL in short). Local-GCL consists of two key designs: 1) We fa bricate the positive examples for each node directly using its first-order neighbors, which frees our method from the reliance on carefully-designed graph augmentations; 2) To improve the efficiency of contrastive learning on graphs, we devise a kernelized contrastive loss, which could be approximately computed in linear time and space complexity with respect to the graph size. We provide theoretical analysis to justify the effectiveness and rationality of the proposed methods. Experiments on various datasets with different scales and properties demonstrate that in spite of its simplicity, Local-GCL achieves quite competitive performance in self-supervised node representation learning tasks on graphs with various scales and properties.

****

OrthoReg: Improving Graph-regularized MLPs via Orthogonality Regularization

Hengrui Zhang,Shen Wang,Soji Adeshina,Vassilis N. Ioannidis,Jiani Zhang,Xiao Qin,Christos Faloutsos,Da Zheng,George Karypis,Philip S. Yu

Graph Neural Networks (GNNs) are currently dominating in modeling graph-structure data, while their high reliance on graph structure for inference significantly impedes them from widespread applications. By contrast, graph-Regularized MLPs (GR-MLPs) implicitly inject the graph structure information into model weights, while their performance can hardly match that of GNNs in most tasks. This motivates us to study the causes of the limited performance of GR-MLPs. In this paper, we demonstrate that node embeddings learned from conventional GR-MLPs suffer from dimensional collapse, a phenomenon in which the largest a few eigenvalues dominate the embedding space, and thus the expressive power is constrained. We further propose ORTHO-REG, a novel GR-MLP model, to mitigate the dimensional collapse issue. Through a soft regularization loss on the correlation matrix of node embeddings, ORTHO-REG explicitly encourages orthogonal node representations and thus can naturally avoid dimensionally collapsed representations. Experiments on traditional transductive semi-supervised classification tasks and inductive node classification for cold-start scenarios demonstrate its effectiveness and superiority.

****

Optimal Activation Functions for the Random Features Regression Model

Jianxin Wang,José Bento

The asymptotic mean squared test error and sensitivity of the Random Features Regression model (RFR) have been recently studied. We build on this work and identify in closed-form the family of Activation Functions (AFs) that minimize a combination of the test error and sensitivity of the RFR under different notions of functional parsimony. We find scenarios under which the optimal AFs are linear, saturated linear functions, or expressible in terms of Hermite polynomials. Finally, we show how using optimal AFs impacts well established properties of the RFR model, such as its double descent curve, and the dependency of its optimal regularization parameter on the observation noise level.

****

Learning to Learn with Generative Models of Neural Network Checkpoints

William Peebles,Ilija Radosavovic,Tim Brooks,Alexei A Efros,Jitendra Malik

We explore a data-driven approach for learning to optimize neural networks. We construct a dataset of neural network checkpoints and train a generative model on the parameters. In particular, our model is a conditional diffusion transformer that, given an initial input parameter vector and a prompted loss, error, or return, predicts the distribution over parameter updates that achieve the desired metric. At test time, it can optimize neural networks with unseen parameters for downstream tasks in just one update. We find that our approach successfully generates parameters for a wide range of loss prompts. Moreover, it can sample multimodal parameter solutions and has favorable scaling properties. We apply our method to different neural network architectures and tasks in supervised and reinforcement learning.

****

Improving Explanation Reliability through Group Attribution

Sangmin Lee,Jongseong Jang,Woohyung Lim

Although input attribution methods are mainstream in understanding predictions of DNNs for straightforward interpretations, the non-linearity of DNNs often makes the attributed scores unreliable in explaining a given prediction, deteriorating the faithfulness of the explanation.

However, the challenge could be mitigated by attributing scores to groups of explanatory components instead of the individuals, termed group attribution. While a group attribution would explain the group-wise contribution more reliably, it does not explain the component-wise contributions so that estimating component-wise scores yields less reliable explanation, indicating the trade-off of group attributions.

In this work, we introduce the generalized definition of reliability loss and group attribution, and formulate the optimization problem of the trade-off with these terms. We apply our formalization to Shapley value attribution to propose the optimization method G-SHAP. We show the effectiveness and explanatory benefits of our method through empirical results on image classification tasks.

****

Intrinsic Motivation via Surprise Memory

Hung Le,Kien Do,Dung Nguyen,Svetha Venkatesh

We present a new computing model for intrinsic rewards in reinforcement learning that addresses the limitations of existing surprise-driven explorations. The reward is the novelty of the surprise rather than the surprise norm. We estimate the surprise novelty as retrieval errors of a memory network wherein the memory stores and reconstructs surprises. Our surprise memory (SM) augments the capability of surprise-based intrinsic motivators, maintaining the agent's interest in exciting exploration while reducing unwanted attraction to unpredictable or noisy observations. Our experiments demonstrate that the SM combined with various surprise predictors exhibits efficient exploring behaviors and significantly boosts the final performance in sparse reward environments, including Noisy-TV, navigation and challenging Atari games.

****

Dr-Fairness: Dynamic Data Ratio Adjustment for Fair Training on Real and Generated Data

Yuji Roh,Weili Nie,De-An Huang,Steven Euijong Whang,Arash Vahdat,Anima Anandkumar

Fair visual recognition has become critical for preventing demographic disparity. A major cause of model unfairness is the imbalanced representation of different groups in training data. Recently, several works aim to alleviate this issue using generated data. However, these approaches often use generated data to obtain similar amounts of data across groups, which is not optimal for achieving high fairness due to different learning difficulties and generated data qualities across groups. To address this issue, we propose a novel adaptive sampling approach that leverages both real and generated data for fairness. We design a bilevel optimization that finds the optimal data sampling ratios among groups and between real and generated data while training a model. The ratios are dynamically adjusted considering both the model's accuracy as well as its fairness. To efficiently solve our non-convex bilevel optimization, we propose a simple approximation to the solution given by the implicit function theorem. Extensive experiments show that our framework achieves state-of-the-art fairness and accuracy on the CelebA and ImageNet People Subtree datasets. We also observe that our method adaptively relies less on the generated data when it has poor quality.

****

Improving Object-centric Learning with Query Optimization

Baoxiong Jia,Yu Liu,Siyuan Huang

The ability to decompose complex natural scenes into meaningful object-centric abstractions lies at the core of human perception and reasoning. In the recent culmination of unsupervised object-centric learning, the Slot-Attention module has played an important role with its simple yet effective design and fostered many powerful variants. These methods, however, have been exceedingly difficult to train
without supervision and are ambiguous in the notion of object, especially for co

mplex natural scenes. In this paper, we propose to address these issues by investigating the potential of learnable queries as initializations for Slot-Attention learning, uniting it with efforts from existing attempts on improving Slot-Attention learning with bi-level optimization. With simple code adjustments on Slot-Attention, our model, Bi-level Optimized Query Slot Attention, achieves state-of-the-art results on 3 challenging synthetic and 7 complex real-world datasets in unsupervised image segmentation and reconstruction, outperforming previous baselines by a large margin. We provide thorough ablative studies to validate the necessity and effectiveness of our design. Additionally, our model exhibits great potential for concept binding and zero-shot learning. Our work is made publicly available at https://bo-qsa.github.io.

****

Set-Level Self-Supervised Learning from Noisily-Labeled Data

Chia-Ching Lin,Shang-Fu Chen,Fu-En Yang,Yu-Chiang Frank Wang,Chin-Laung Lei

Noisy labels are inevitably presented in real-world datasets due to labeling error or visual content ambiguity. Existing methods generally approach the task of noisy label learning (NLL) by either properly regularizing the model, or reweighting clean/noisy labeled samples. While self-supervised learning (SSL) has been applied to pre-train deep neural networks without label supervision, downstream tasks like image classification still require clean labeled data. And, most SSL strategies are performed at the instance level, regardless of the correctness of its label. In this paper, we propose set-level self-supervised learning (SLSSL), which performs SSL at mini-batch levels with observed noisy labels. By corrupting the labels of each training mini-batch, our SLSSL enforces the model to exhibit sufficient robustness. Moreover, the proposed SLSSL can also be utilized for sample reweighting technique. As a result, the proposed learning scheme can be applied as an expectation-maximization (EM) algorithm during model training. Extensive experiments on synthetic and real-world noisy label data confirm the effectiveness of our framework.

****

Feature Reconstruction From Outputs Can Mitigate Simplicity Bias in Neural Networks

Sravanti Addepalli,Anshul Nasery,Venkatesh Babu Radhakrishnan,Praneeth Netrapalli,Prateek Jain

Deep Neural Networks are known to be brittle to even minor distribution shifts compared to the training distribution. While one line of work has demonstrated that \emph{Simplicity Bias} (SB) of DNNs -- bias towards learning only the simplest features -- is a key reason for this brittleness, another recent line of work has surprisingly found that diverse/ complex features are indeed learned by the backbone, and their brittleness is due to the linear classification head relying primarily on the simplest features. To bridge the gap between these two lines of work, we first hypothesize and verify that while SB may not altogether preclude learning complex features, it amplifies simpler features over complex ones. Namely, simple features are replicated several times in the learned representations while complex features might not be replicated. This phenomenon, we term \emph{Feature Replication Hypothesis}, coupled with the \emph{Implicit Bias} of SGD to converge to maximum margin solutions in the feature space, leads the models to rely mostly on the simple features for classification. To mitigate this bias, we propose \emph{Feature Reconstruction Regularizer (FRR)} to ensure that the learned features can be reconstructed back from the logits. The use of \emph{FRR} in linear layer training (\emph{FRR-L}) encourages the use of more diverse features for classification. We further propose to finetune the full network by freezing the weights of the linear layer trained using \emph{FRR-L}, to refine the learned features, making them more suitable for classification. Using this simple solution, we demonstrate up to 15\% gains in OOD accuracy on the recently introduced semi-synthetic datasets with extreme distribution shifts. Moreover, we demonstrate noteworthy gains over existing SOTA methods on the standard OOD benchmark DomainBed as well.

****

EUCLID: Towards Efficient Unsupervised Reinforcement Learning with Multi-choice

Dynamics Model

Yifu Yuan,Jianye HAO,Fei Ni,Yao Mu,YAN ZHENG,Yujing Hu,Jinyi Liu,Yingfeng Chen,Changjie Fan

Unsupervised reinforcement learning (URL) poses a promising paradigm to learn useful behaviors in a task-agnostic environment without the guidance of extrinsic rewards to facilitate the fast adaptation of various downstream tasks. Previous works focused on the pre-training in a model-free manner while lacking the study of transition dynamics modeling that leaves a large space for the improvement of sample efficiency in downstream tasks. To this end, we propose an Efficient Unsupervised Reinforcement Learning Framework with Multi-choice Dynamics model (EUCLID), which introduces a novel model-fused paradigm to jointly pre-train the dynamics model and unsupervised exploration policy in the pre-training phase, thus better leveraging the environmental samples and improving the downstream task sampling efficiency. However, constructing a generalizable model which captures the local dynamics under different behaviors remains a challenging problem. We introduce the multi-choice dynamics model that covers different local dynamics under different behaviors concurrently, which uses different heads to learn the state transition under different behaviors during unsupervised pre-training and selects the most appropriate head for prediction in the downstream task. Experimental results in the manipulation and locomotion domains demonstrate that EUCLID achieves state-of-the-art performance with high sample efficiency, basically solving the state-based URLB benchmark and reaching a mean normalized score of 104.0±1.2% in downstream tasks with 100k fine-tuning steps, which is equivalent to DDPG’s performance at 2M interactive steps with 20× more data. More visualization videos are released on our homepage.

****

A General Framework for Sample-Efficient Function Approximation in Reinforcement Learning

Zixiang Chen,Chris Junchi Li,Huizhuo Yuan,Quanquan Gu,Michael Jordan

With the increasing need for handling large state and action spaces, general function approximation has become a key technique in reinforcement learning (RL). In this paper, we propose a general framework that unifies model-based and model-free RL, and an Admissible Bellman Characterization (ABC) class that subsumes nearly all Markov decision process (MDP) models in the literature for tractable RL. We propose a novel estimation function with decomposable structural properties for optimization-based exploration and the functional Eluder dimension as a complexity measure of the ABC class. Under our framework, a new sample-efficient algorithm namely OPtimization-based ExploRation with Approximation (OPERA) is proposed, achieving regret bounds that match or improve over the best-known results for a variety of MDP models. In particular, for MDPs with low Witness rank, under a slightly stronger assumption, OPERA improves the state-of-the-art sample complexity results by a factor of dH. Our framework provides a generic interface to design and analyze new RL models and algorithms.

****

Maximizing Spatio-Temporal Entropy of Deep 3D CNNs for Efficient Video Recognition

Junyan Wang,Zhenhong Sun,Yichen Qian,Dong Gong,Xiuyu Sun,Ming Lin,Maurice Pagnucco,Yang Song

3D convolution neural networks (CNNs) have been the prevailing option for video recognition. To capture the temporal information, 3D convolutions are computed along the sequences, leading to cubically growing and expensive computations. To reduce the computational cost, previous methods resort to manually designed 3D/2D CNN structures with approximations or automatic search, which sacrifice the modeling ability or make training time-consuming. In this work, we propose to automatically design efficient 3D CNN architectures via a novel training-free neural architecture search approach tailored for 3D CNNs considering the model complexity. To measure the expressiveness of 3D CNNs efficiently, we formulate a 3D CNN as an information system and derive an analytic entropy score, based on the Maximum Entropy Principle. Specifically, we propose a spatio-temporal entropy score (STEntr-Score) with a refinement factor to handle the discrepancy of visual inf ormation in spatial and temporal dimensions, through dynamically leveraging the correlation between the feature map size and kernel size depth-wisely. Highly efficient and expressive 3D CNN architectures, i.e., entropy-based 3D CNNs (E3D family), can then be efficiently searched by maximizing the STEntr-Score under a given computational budget, via an evolutionary algorithm without training the network parameters. Extensive experiments on Something-Something V1&V2 and Kinetics400 demonstrate that the E3D family achieves state-of-the-art performance with higher computational efficiency.

****

Cycle to Clique (Cy2C) Graph Neural Network: A Sight to See beyond Neighborhood Aggregation

Yun Young Choi,Sun Woo Park,Youngho Woo,U Jin Choi

Graph neural networks have been successfully adapted for learning vector representations of graphs through various neighborhood aggregation schemes. Previous researches suggest, however, that they possess limitations in incorporating key non-Euclidean topological properties of graphs. This paper mathematically identifies the caliber of graph neural networks in classifying isomorphism classes of graphs with continuous node attributes up to their local topological properties. In light of these observations, we construct the Cycle to Clique graph neural network, a novel yet simple algorithm which topologically enriches the input data of conventional graph neural networks while preserving their architectural components. This method theoretically outperforms conventional graph neural networks in classifying isomorphism classes of graphs while ensuring comparable time complexity in representing random graphs. Empirical results further support that the novel algorithm produces comparable or enhanced results in classifying benchmark graph data sets compared to contemporary variants of graph neural networks.

****

Latent State Marginalization as a Low-cost Approach for Improving Exploration

Dinghuai Zhang,Aaron Courville,Yoshua Bengio,Qinqing Zheng,Amy Zhang,Ricky T. Q. Chen

While the maximum entropy (MaxEnt) reinforcement learning (RL) framework -- often touted for its exploration and robustness capabilities -- is usually motivated from a probabilistic perspective, the use of deep probabilistic models have not gained much traction in practice due to their inherent complexity. In this work, we propose the adoption of latent variable policies within the MaxEnt framework, which we can provably approximate any policy distribution, and additionally, naturally emerges under the use of world models with a latent belief state.

We discuss why latent variable policies are difficult to train, how naive approaches can fail, and subsequently introduce a series of improvements centered around low-cost marginalization of the latent state, allowing us to make full use of the latent state at minimal additional cost. We instantiate our method under the actor-critic framework, marginalizing both the actor and critic. The resulting algorithm, referred to as Stochastic Marginal Actor-Critic (SMAC), is simple yet effective. We experimentally validate our method on continuous control tasks, showing that effective marginalization can lead to better exploration and more robust training. Our implementation is open sourced at https://github.com/zdhNarsil/Stochastic-Marginal-Actor-Critic.

****

TensorVAE: A Direct Generative Model for Molecular Conformation Generation driven by Novel Feature Engineering

Hongyang Yu,Hongjiang Yu

Efficient generation of 3D conformations of a molecule from its 2D graph is a key challenge in in-silico drug discovery. Deep learning (DL) based generative modelling has recently become a potent tool to tackling this challenge. However, many existing DL-based methods are either indirect-leveraging inter-atomic distances or direct-but requiring complex feature transformation or numerous sampling steps to generate conformations. In this work, we propose a simple model abbreviated TensorVAE capable of generating conformations directly from a 2D molecular graph in a single step. The main novelty of the proposed method is focused on feature engineering. We develop a novel encoding and feature extraction mechanism r elying solely on standard convolution operation to generate token-like feature vector for each atom. These feature vectors are then transformed through standard transformer encoders under a conditional Variational Auto Encoder framework for learning to generate conformations directly. We show through experiments on two benchmark datasets that with intuitive and sensible feature engineering, a relatively simple and standard model can provide promising generative capability rivalling recent state-of-the-art models employing more sophisticated and specialized generative architecture.

****

Smoothed-SGDmax: A Stability-Inspired Algorithm to Improve Adversarial Generalization

Jiancong Xiao,Jiawei Zhang,Zhi-Quan Luo,Asuman E. Ozdaglar

Unlike standard training, deep neural networks can suffer from serious overfitting problems in adversarial settings, which is studied extensively by empirical papers. Recent research (e.g., Xing et al. (2021); Xiao et al. (2022)) show that SGDmax-based adversarial training algorithms with $1/s(T)$ training loss incurs a stability-based generalization bound in $\Theta(c+s(T)/n)$. Here T is the number of iterations, n is the number of samples, $s(T)\rightarrow \infty$ as $T\rightarrow \infty$, and c is a n-independent term. This reveals that adversarial training can have nonvanishing generalization errors even if the sample size n goes to infinity. A natural question arises: can we eliminate the nonvanishing term c by designing a more generalizable algorithm? We give an affirmative answer in this paper. First, by an adaptation of information-theoretical lower bound on the complexity of solving Lipschitz-convex problems using randomized algorithms, we show that a minimax lower bound for adversarial generalization gap is $\Omega(s(T)/n)$ given training loss $1/s(T)$. This implies that SGDmax does not achieve the lower bound. Next, by observing that the nonvanishing generalization error term for SGDmax comes from the non-smoothness of the adversarial loss function, we employ a smoothing technique to smooth the adversarial loss function. Based on the smoothed loss function, we design a smoothed SGDmax algorithm achieving generalization bound $\mathcal{O}(s(T)/n)$, which matches the minimax lower bound. Experimentally, we show that our algorithm improves adversarial generalization on common datasets.

****

Generalizing and Decoupling Neural Collapse via Hyperspherical Uniformity Gap

Weiyang Liu,Longhui Yu,Adrian Weller,Bernhard Schölkopf

The neural collapse (NC) phenomenon describes an underlying geometric symmetry for deep neural networks, where both deeply learned features and classifiers converge to a simplex equiangular tight frame. It has been shown that both cross-entropy loss and mean square error can provably lead to NC. We remove NC's key assumption on the feature dimension and the number of classes, and then present a generalized neural collapse (GNC) hypothesis that effectively subsumes the original NC. Inspired by how NC characterizes the training target of neural networks, we decouple GNC into two objectives: minimal intra-class variability and maximal inter-class separability. We then use hyperspherical uniformity (which characterizes the degree of uniformity on the unit hypersphere) as a unified framework to quantify these two objectives. Finally, we propose a general objective -- hyperspherical uniformity gap (HUG), which is defined by the difference between inter-class and intra-class hyperspherical uniformity. HUG not only provably converges to GNC, but also decouples GNC into two separate objectives. Unlike cross-entropy loss that couples intra-class compactness and inter-class separability, HUG enjoys more flexibility and serves as a good alternative loss function. Empirical results show that HUG works well in terms of generalization and robustness.

****

MaskFusion: Feature Augmentation for Click-Through Rate Prediction via Input-adaptive Mask Fusion

Chao Liao,Jianchao Tan,Jiyuan Jia,Yi Guo,Chengru Song

Click-through rate (CTR) prediction plays important role in the advertisement, recommendation, and retrieval applications. Given the feature set, how to fully utilize the information from the feature set is an active topic in deep CTR model designs. There are several existing deep CTR works focusing on feature interactions, feature attentions, and so on. They attempt to capture high-order feature interactions to enhance the generalization ability of deep CTR models. However, these works either suffer from poor high-order feature interaction modeling using DNN or ignore the balance between generalization and memorization during the recommendation. To mitigate these problems, we propose an adaptive feature fusion framework called MaskFusion, to additionally capture the explicit interactions between the input feature and the existing deep part structure of deep CTR models dynamically, besides the common feature interactions proposed in existing works. MaskFusion is an instance-aware feature augmentation method, which makes deep CTR models more personalized by assigning each feature with an instance-adaptive mask and fusing each feature with each hidden state vector in the deep part structure. MaskFusion can also be integrated into any existing deep CTR models flexibly. MaskFusion achieves state-of-the-art (SOTA) performance on all seven benchmarks deep CTR models with three public datasets.

****

Finite-time Analysis of Single-timescale Actor-Critic on Linear Quadratic Regulator

Xuyang Chen,Jingliang Duan,Lin Zhao

Actor-critic (AC) methods have achieved state-of-the-art performance in many challenging tasks. However, their convergence in most practical applications are still poorly understood. Existing works mostly consider the uncommon double-loop or two-timescale stepsize variants for the ease of analysis. We investigate the practical yet more challenging vanilla single-sample single-timescale AC for solving the canonical linear quadratic regulator problem. Specifically, the actor and the critic update only once with a single sample in each iteration using proportional stepsizes. We prove that the vanilla AC can attain an ϵ-optimal solution with a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$, which elucidates on the practical efficiency of single-sample single-timescale AC. We develop a novel analysis framework that directly bounds the whole interconnected iteration system without the conservative decoupling commonly adopted in previous analysis of AC. Our work presents the first finite-time analysis of single-sample single-timescale AC with a global optimality guarantee.

****

Scalable 3D Object-centric Learning

Tianyu Wang,miaomiao Liu,Kee Siong Ng

We tackle the task of unsupervised 3D object-centric representation learning on scenes of potentially unbounded scale.

Existing approaches to object-centric representation learning exhibit significant limitations in achieving scalable inference due to their dependencies on a fixed global coordinate system.

In contrast, we propose to learn view-invariant 3D object representations in localized object coordinate systems.

To this end, we estimate the object pose and appearance representation separately and explicitly project object representations across views.

We adopt amortized variational inference to process sequential input and update object representations online.

To scale up our model to scenes with an arbitrary number of objects, we further introduce a Cognitive Map that allows the registration and querying of objects on a global map.

We employ the object-centric neural radiance field (NeRF) as our 3D scene representation, which is jointly inferred by our unsupervised object-centric learning framework.

Experimental results demonstrate that our method can infer and maintain object-centric representations of unbounded 3D scenes.

Further combined with a per-object NeRF finetuning process, our model can achieve scalable high-quality object-aware scene reconstruction.

****

Towards Boosting the Open-Domain Chatbot with Human Feedback

Hua Lu,Siqi Bao,Huang He,Fan Wang,Hua Wu,Haifeng Wang

Many open-domain dialogue models pre-trained with social media comments can generate coherent replies but have difficulties producing engaging responses. This phenomenon might mainly result from the deficiency of annotated human-human conversations and the misalignment with human preference. In this paper, we propose a novel and efficient framework Diamante to boost the open-domain chatbot, where two kinds of human feedback (including explicit demonstration and implicit preference) are collected and leveraged. By asking annotators to select or amend the model-generated candidate responses, Diamante efficiently collects the human demonstrated responses and constructs a Chinese chit-chat dataset. To enhance the alignment with human preference, Diamante leverages the implicit preference in the data collection process and introduces the generation-evaluation joint training. Comprehensive experiments indicate that the Diamante dataset and joint training paradigm can significantly boost the performance of pre-trained dialogue models. The overall engagingness of the previous state-of-the-art model has been improved remarkably by 50% in Chinese open-domain conversations.

****

Learning to Generate All Feasible Actions

Mirco Theile,Daniele Bernardini,Raphael Trumpp,Cristina Piazza,Marco Caccamo,Alberto Sangiovanni-Vincentelli

Several machine learning (ML) applications are characterized by searching for an optimal solution to a complex task. The search space for this optimal solution is often very large, so large in fact that this optimal solution is often not computable. Part of the problem is that many candidate solutions found via ML are actually infeasible and have to be discarded. Restricting the search space to only the feasible solution candidates simplifies finding an optimal solution for the tasks. Further, the set of feasible solutions could be re-used in multiple problems characterized by different tasks. In particular, we observe that complex tasks can be decomposed into subtasks and corresponding skills. We propose to learn a reusable and transferable skill by training an actor to generate all feasible actions. The trained actor can then propose feasible actions, among which an optimal one can be chosen according to a specific task. The actor is trained by interpreting the feasibility of each action as a target distribution. The training procedure minimizes a divergence of the actor's output distribution to this target. We derive the general optimization target for arbitrary f-divergences using a combination of kernel density estimates, resampling, and importance sampling. We further utilize an auxiliary critic to reduce the interactions with the environment. A preliminary comparison to related strategies shows that our approach learns to visit all the modes in the feasible action space, demonstrating the framework's potential for generating multimodal action distributions.

****

Rethinking Self-Supervised Visual Representation Learning in Pre-training for 3D Human Pose and Shape Estimation

Hongsuk Choi,Hyeongjin Nam,Taeryung Lee,Gyeongsik Moon,Kyoung Mu Lee

Recently, a few self-supervised representation learning (SSL) methods have outperformed the ImageNet classification pre-training for vision tasks such as object detection. However, its effects on 3D human body pose and shape estimation (3DHPSE) are open to question, whose target is fixed to a unique class, the human, and has an inherent task gap with SSL. We empirically study and analyze the effects of SSL and further compare it with other pre-training alternatives for 3DHPSE. The alternatives are 2D annotation-based pre-training and synthetic data pre-training, which share the motivation of SSL that aims to reduce the labeling cost. They have been widely utilized as a source of weak-supervision or fine-tuning, but have not been remarked as a pre-training source. SSL methods underperform the conventional ImageNet classification pre-training on multiple 3DHPSE benchmarks by 7.7% on average. In contrast, despite a much less amount of pre-training data, the 2D annotation-based pre-training improves accuracy on all benchmarks and shows faster convergence during fine-tuning. Our observations challenge the naive application of the current SSL pre-training to 3DHPSE and relight the value of other data types in the pre-training aspect.

****

Sparsity by Redundancy: Solving L_1 with a Simple Reparametrization
Liu Ziyin,Zihao Wang

We identify and prove a general principle: L_1 sparsity can be achieved using a redundant parametrization plus L_2 penalty. Our results lead to a simple algorithm, \textit{spred}, that seamlessly integrates L_1 regularization into any modern deep learning framework. Practically, we demonstrate (1) the efficiency of \textit{spred} in optimizing conventional tasks such as lasso and sparse coding, (2) benchmark our method for nonlinear feature selection of six gene selection tasks, and (3) illustrate the usage of the method for achieving structured and unstructured sparsity in deep learning in an end-to-end manner. Conceptually, our result bridges the gap in understanding the inductive bias of the redundant parametrization common in deep learning and conventional statistical learning.

****

Test-Time Adaptation for Visual Document Understanding
Sayna Ebrahimi,Sercan O Arik,Tomas Pfister

Self-supervised pretraining has been able to produce transferable representations for various visual document understanding (VDU) tasks. However, the ability of such representations to adapt to new distribution shifts at test-time has not been studied yet. We propose DocTTA, a novel test-time adaptation approach for documents that leverages cross-modality self-supervised learning via masked visual language modeling as well as pseudo labeling to adapt models learned on a \textit{source} domain to an unlabeled \textit{target} domain at test time. We also introduce new benchmarks using existing public datasets for various VDU tasks including entity recognition, key-value extraction, and document visual question answering tasks where DocTTA improves the source model performance up to 1.79\% in (F1 score), 3.43\% (F1 score), and 17.68\% (ANLS score), respectively.

****

Learned Index with Dynamic ϵ
Daoyuan Chen,Wuchao Li,Yaliang Li,Bolin Ding,Kai Zeng,Defu Lian,Jingren Zhou

Index structure is a fundamental component in database and facilitates broad data retrieval applications. Recent learned index methods show superior performance by learning hidden yet useful data distribution with the help of machine learning, and provide a guarantee that the prediction error is no more than a pre-defined ϵ. However, existing learned index methods adopt a fixed ϵ for all the learned segments, neglecting the diverse characteristics of different data localities. In this paper, we propose a mathematically-grounded learned index framework with dynamic ϵ, which is efficient and pluggable to existing learned index methods. We theoretically analyze prediction error bounds that link ϵ with data characteristics for an illustrative learned index method. Under the guidance of the derived bounds, we learn how to vary ϵ and improve the index performance with a better space-time trade-off. Experiments with real-world datasets and several state-of-the-art methods demonstrate the efficiency, effectiveness and usability of the proposed framework.

****

Boosting Multiagent Reinforcement Learning via Permutation Invariant and Permutation Equivariant Networks
Jianye HAO,Xiaotian Hao,Hangyu Mao,Weixun Wang,Yaodong Yang,Dong Li,YAN ZHENG,Zhen Wang

The state space in Multiagent Reinforcement Learning (MARL) grows exponentially with the agent number. Such a curse of dimensionality results in poor scalability and low sample efficiency, inhibiting MARL for decades. To break this curse, we propose a unified agent permutation framework that exploits the permutation invariance (PI) and permutation equivariance (PE) inductive biases to reduce the multiagent state space. Our insight is that permuting the order of entities in the factored multiagent state space does not change the information. Specifically, we propose two novel implementations: a Dynamic Permutation Network (DPN) and a Hyper Policy Network (HPN). The core idea is to build separate entity-wise PI input and PE output network modules to connect the entity-factored state space and action space in an end-to-end way. DPN achieves such connections by two separa te module selection networks, which consistently assign the same input module to the same input entity (guarantee PI) and assign the same output module to the same entity-related output (guarantee PE). To enhance the representation capability, HPN replaces the module selection networks of DPN with hypernetworks to directly generate the corresponding module weights. Extensive experiments in SMAC, Google Research Football and MPE validate that the proposed methods significantly boost the performance and the learning efficiency of existing MARL algorithms. Remarkably, in SMAC, we achieve 100% win rates in almost all hard and super-hard scenarios (never achieved before).

****

LAU: A novel two-parameter learnable Logmoid Activation Unit

Xue-Mei Zhou,Ling-Fang Li,Xing-Zhou Zheng,Mingxing Luo

The activation function in deep neural networks has a major impact on the performance of the training stage. In this work, we proposed a novel learnable Logmoid Activation Unit (LAU), $f(x)=x\ln(1+\alpha \textrm{sigmoid}(\beta x))$, with two free parameters α and β that can be optimized via back-propagation algorithm. We design quasi-interpolation type neural network operators with Logmoid-1 in a given feed-forward neural network for approximating any continuous function in closed spaces. This provides a theoretical basis for the excellent empirical performance of LAUs in experimental simulations. For instance, compared with ReLUs the proposed LAUs improves Top-1 classification accuracy on ImageNet-200 by 7% respectively in ShuffleNet-V2, on CIFAR-10 by 6$\%$ respectively in EfficientNet-B0, and on CIFAR-100 by 5$\%$ respectively in MobileNet-V2. Our simulations show that end-to-end learning deep neural networks with learnable Logmoids can increase the predictive performance.

****

N-Student Learning: An Approach to Model Uncertainty and Combat Overfitting

Ryan Y. Xu,Solomon E Garber,Antonella Di Lillo,James Storer

This work presents N-Student Learning, a pseudo-label based multi-network training setup that can be applied to nearly any supervised learning architecture in order to help combat the problem of overfitting and control the way in which a network models uncertainty in the data. The effectiveness of N-Student Learning relies on the idea that a network's predictions on unseen data are largely independent of any instance-dependent noise in the labels. In N-Student Learning, each student network is assigned a subset of the training dataset such that no data point is in every student's training subset. Unbiased pseudo-labels can thus be generated for every data point in the training set by taking the predictions of appropriate student networks. Training on these unbiased pseudo-labels minimizes the extent to which each network overfits to instance-dependent noise in the data. Furthermore, based on prior knowledge of the domain, we can control how the networks learn to model uncertainty that is present in the dataset by adjusting the way that pseudo-labels are generated. While this method is largely inspired by the general problem of overfitting, a natural application is found in the problem of classification with noisy labels — a domain where overfitting is a significant concern. After developing intuition through a toy classification task, we proceed to demonstrate that N-Student Learning performs favorably on benchmark datasets when compared to state-of-the-art methods in the problem of classification with noisy labels.

****

wav2tok: Deep Sequence Tokenizer for Audio Retrieval

Adhiraj Banerjee,Vipul Arora

Search over audio sequences is a fundamental problem. In this paper, we propose a method to extract concise discrete representations for audio that can be used for efficient retrieval. Our motivation comes from orthography which represents speech of a given language in a concise and distinct discrete form. The proposed method, wav2tok, learns such representations for any kind of audio, speech or non-speech, from pairs of similar audio. wav2tok compresses the query and target sequences into shorter sequences of tokens that are faster to match. The learning method makes use of CTC loss and expectation-maximization algorithm, which are generally used for supervised automatic speech recognition and for learning dis crete latent variables, respectively. Experiments show the consistent performance of wav2tok across two audio retrieval tasks: music search (query by humming) and speech search via audio query, outperforming state-of-the-art baselines.

****

Image to Sphere: Learning Equivariant Features for Efficient Pose Prediction
David Klee,Ondrej Biza,Robert Platt,Robin Walters

Predicting the pose of objects from a single image is an important but difficult computer vision problem. Methods that predict a single point estimate do not predict the pose of objects with symmetries well and cannot represent uncertainty. Alternatively, some works predict a distribution over orientations in $\mathrm{SO}(3)$. However, training such models can be computation- and sample-inefficient. Instead, we propose a novel mapping of features from the image domain to the 3D rotation manifold. Our method then leverages $\mathrm{SO}(3)$ equivariant layers, which are more sample efficient, and outputs a distribution over rotations that can be sampled at arbitrary resolution. We demonstrate the effectiveness of our method at object orientation prediction, and achieve state-of-the-art performance on the popular PASCAL3D+ dataset. Moreover, we show that our method can model complex object symmetries, without any modifications to the parameters or loss function. Code is available at \url{https://dmklee.github.io/image2sphere}.

****

PV3D: A 3D Generative Model for Portrait Video Generation
Eric Zhongcong Xu,Jianfeng Zhang,Jun Hao Liew,Wenqing Zhang,Song Bai,Jiashi Feng,Mike Zheng Shou

Recent advances in generative adversarial networks (GANs) have demonstrated the capabilities of generating stunning photo-realistic portrait images. While some prior works have applied such image GANs to unconditional 2D portrait video generation and static 3D portrait synthesis, there are few works successfully extending GANs for generating 3D-aware portrait videos. In this work, we propose PV3D, the first generative framework that can synthesize multi-view consistent portrait videos. Specifically, our method extends the recent static 3D-aware image GAN to the video domain by generalizing the 3D implicit neural representation to model the spatio-temporal space. To introduce motion dynamics into the generation process, we develop a motion generator by stacking multiple motion layers to generate motion features via modulated convolution. To alleviate motion ambiguities caused by camera/human motions, we propose a simple yet effective camera condition strategy for PV3D, enabling both temporal and multi-view consistent video generation. Moreover, PV3D introduces two discriminators for regularizing the spatial and temporal domains to ensure the plausibility of the generated portrait videos. These elaborated designs enable PV3D to generate 3D-aware motion-plausible portrait videos with high-quality appearance and geometry, significantly outperforming prior works. As a result, PV3D is able to support downstream applications such as static portrait animation and view-consistent motion editing. Code and models are available at https://showlab.github.io/pv3d.

****

k-Median Clustering via Metric Embedding: Towards Better Initialization with Differential Privacy
Chenglin Fan,Ping Li,Xiaoyun Li

In clustering algorithms, the choice of initial centers is crucial for the quality of the learned clusters. We propose a new initialization scheme for the k-median problem in the general metric space (e.g., discrete space induced by graphs), based on the construction of metric embedding tree structure of the data. From the tree, we propose a novel and efficient search algorithm, for good initial centers that can be used subsequently for the local search algorithm. The so-called HST initialization method can produce initial centers achieving lower errors than those from another popular initialization method, k-median++, with comparable efficiency. Our HST initialization can also be easily extended to the setting of differential privacy (DP) to generate private initial centers. We show that the error of applying DP local search followed by our private HST initialization improves previous results on the approximation error, and approaches the lower bound within a small factor. Experiments demonstrate the effectiveness of ou r proposed methods.

****

Analysis of Error Feedback in Compressed Federated Non-Convex Optimization

Xiaoyun Li,Ping Li

Communication cost between the clients and the central server could be a bottleneck in real-world Federated Learning (FL) systems. In classical distributed learning, the method of Error Feedback (EF) has been a popular technique to remedy the downsides of biased gradient compression, but literature on applying EF to FL is still very limited. In this work, we propose a compressed FL scheme equipped with error feedback, named Fed-EF, with two variants depending on the global optimizer. We provide theoretical analysis showing that Fed-EF matches the convergence rate of the full-precision FL counterparts in non-convex optimization under data heterogeneity. Moreover, we initiate the first analysis of EF under partial client participation, which is an important scenario in FL, and demonstrate that the convergence rate of Fed-EF exhibits an extra slow down factor due to the ``stale error compensation'' effect. Experiments are conducted to validate the efficacy of Fed-EF in practical FL tasks and justify our theoretical findings.

****

Characterizing the Influence of Graph Elements

Zizhang Chen,Peizhao Li,Hongfu Liu,Pengyu Hong

Influence function, a method from the robust statistics, measures the changes of model parameters or some functions about model parameters with respect to the removal or modification of training instances. It is an efficient and useful post-hoc method for studying the interpretability of machine learning models without the need of expensive model re-training. Recently, graph convolution networks (GCNs), which operate on graph data, have attracted a great deal of attention. However, there is no preceding research on the influence functions of GCNs to shed light on the effects of removing training nodes/edges from an input graph. Since the nodes/edges in a graph are interdependent in GCNs, it is challenging to derive influence functions for GCNs. To fill this gap, we started with the simple graph convolution (SGC) model that operates on an attributed graph, and formulated an influence function to approximate the changes of model parameters when a node or an edge is removed from an attributed graph. Moreover, we theoretically analyzed the error bound of the estimated influence of removing an edge. We experimentally validated the accuracy and effectiveness of our influence estimation function. In addition, we showed that the influence function of a SGC model could be used to estimate the impact of removing training nodes/edges on the test performance of the SGC without re-training the model. Finally, we demonstrated how to use influence functions to effectively guide the adversarial attacks on GCNs.

****

Adversarially Robust Neural Lyapunov Control

Li Wei,Yuankun Jiang,Chenglin Li,Wenrui Dai,Junni Zou,Hongkai Xiong

State-of-the-art learning-based stability control methods for nonlinear robotic systems suffer from the issue of reality gap, which stems from discrepancy of the system dynamics between training and target (test) environments. To mitigate this gap, we propose an adversarially robust neural Lyapunov control (ARNLC) method to improve the robustness and generalization capabilities for Lyapunov theory-based stability control. Specifically, inspired by adversarial learning, we introduce an adversary to simulate the dynamics discrepancy, which is learned through deep reinforcement learning to generate the worst-case perturbations during the controller's training. By alternatively updating the controller to minimize the perturbed Lyapunov risk and the adversary to deviate the controller from its objective, the learned control policy enjoys a theoretical guarantee of stability. Empirical evaluations on five stability control tasks with the uniform and worst-case perturbations demonstrate that ARNLC not only accelerates the convergence to asymptotic stability, but can generalize better in the entire perturbation space.

****

MICN: Multi-scale Local and Global Context Modeling for Long-term Series Forecasting

Huiqiang Wang,Jian Peng,Feihu Huang,Jince Wang,Junhui Chen,Yifei Xiao

Recently, Transformer-based methods have achieved surprising performance in the field of long-term series forecasting, but the attention mechanism for computing global correlations entails high complexity. And they do not allow for targeted modeling of local features as CNN structures do. To solve the above problems, we propose to combine local features and global correlations to capture the overall view of time series (e.g., fluctuations, trends). To fully exploit the underlying information in the time series, a multi-scale branch structure is adopted to model different potential patterns separately. Each pattern is extracted with down-sampled convolution and isometric convolution for local features and global correlations, respectively. In addition to being more effective, our proposed method, termed as Multi-scale Isometric Convolution Network (MICN), is more efficient with linear complexity about the sequence length with suitable convolution kernels. Our experiments on six benchmark datasets show that compared with state-of-the-art methods, MICN yields 17.2% and 21.6% relative improvements for multivariate and univariate time series, respectively.

****

EMP: Effective Multidimensional Persistence for Graph Representation Learning

Ignacio Segovia-Dominguez,Baris Coskunuzer,Cuneyt Gurcan Akcora,Yuzhou Chen,Zhiwei Zhen,Murat Kantarcioglu,Yulia Gel

Topological data analysis (TDA) has become increasingly popular in a broad range of machine learning tasks, ranging from anomaly detection and manifold learning to graph classification.

Persistent homology being the key approach in TDA provides a unique topological fingerprint of the data by assessing the evolution of various hidden patterns in the data as we vary a scale parameter. Current PH tools are limited to analyze the data by filtering with single parameter while in many applications, several relevant parameters are equally important to get a much finer information on the data. In this paper, we overcome this problem by introducing Effective Multidimensional Persistence (EMP) framework which enables to investigate the data by varying multiple scale parameters simultaneously. EMP framework provides a highly expressive summary of the data by integrating the multiple descriptor functions to the process successfully. EMP naturally adapts to many known single PH summaries and convert them into multidimensional summaries, for example, EMP Landscapes, EMP Silhouettes, EMP Images, and EMP Surfaces. These summaries deliver a multidimensional fingerprint of the data as matrices and arrays which are suitable for various machine learning models.

We apply EMP framework in graph classification tasks and observe that EMP boosts the performances of various single PH descriptors, and outperforms the most state-of-the-art methods on benchmark datasets. We further derive theoretical guarantees of the proposed EMP summary and prove the stability properties.

****

SWIFT: Rapid Decentralized Federated Learning via Wait-Free Model Communication

Marco Bornstein,Tahseen Rabbani,Evan Z Wang,Amrit Bedi,Furong Huang

The decentralized Federated Learning (FL) setting avoids the role of a potentially unreliable or untrustworthy central host by utilizing groups of clients to collaboratively train a model via localized training and model/gradient sharing. Most existing decentralized FL algorithms require synchronization of client models where the speed of synchronization depends upon the slowest client. In this work, we propose SWIFT: a novel wait-free decentralized FL algorithm that allows clients to conduct training at their own speed. Theoretically, we prove that SWIFT matches the gold-standard iteration convergence rate $\mathcal{O}(1/\sqrt{T})$ of parallel stochastic gradient descent for convex and non-convex smooth optimization (total iterations T). Furthermore, we provide theoretical results for IID and non-IID settings without any bounded-delay assumption for slow clients which is required by other asynchronous decentralized FL algorithms. Although SWIFT achieves the same iteration convergence rate with respect to T as other state-of-the-art (SOTA) parallel stochastic algorithms, it converges faster with respect to runtime due to its wait-free structure. Our experimental results demonstrate that SWIFT's runtime is reduced due to a large reduction in communication t ime per epoch, which falls by an order of magnitude compared to synchronous counterparts. Furthermore, SWIFT produces loss levels for image classification, over IID and non-IID data settings, upwards of 50\% faster than existing SOTA algorithms.

****

Hierarchical Sliced Wasserstein Distance

Khai Nguyen,Tongzheng Ren,Huy Nguyen,Litu Rout,Tan Minh Nguyen,Nhat Ho

Sliced Wasserstein (SW) distance has been widely used in different application scenarios since it can be scaled to a large number of supports without suffering from the curse of dimensionality. The value of sliced Wasserstein distance is the average of transportation cost between one-dimensional representations (projections) of original measures that are obtained by Radon Transform (RT). Despite its efficiency in the number of supports, estimating the sliced Wasserstein requires a relatively large number of projections in high-dimensional settings. Therefore, for applications where the number of supports is relatively small compared with the dimension, e.g., several deep learning applications where the mini-batch approaches are utilized, the complexities from matrix multiplication of Radon Transform become the main computational bottleneck. To address this issue, we propose to derive projections by linearly and randomly combining a smaller number of projections which are named bottleneck projections. We explain the usage of these projections by introducing Hierarchical Radon Transform (HRT) which is constructed by applying Radon Transform variants recursively. We then formulate the approach into a new metric between measures, named Hierarchical Sliced Wasserstein (HSW) distance. By proving the injectivity of HRT, we derive the metricity of HSW. Moreover, we investigate the theoretical properties of HSW including its connection to SW variants and its computational and sample complexities. Finally, we compare the computational cost and generative quality of HSW with the conventional SW on the task of deep generative modeling using various benchmark datasets including CIFAR10, CelebA, and Tiny ImageNet.

****

Test-time Adaptation for Better Adversarial Robustness

Zhichao Huang,Chen Liu,Mathieu Salzmann,Sabine Süsstrunk,Tong Zhang

Standard adversarial training and its variants have been widely adopted in practice to achieve robustness against adversarial attacks. However, we show in this work that such an approach does not necessarily achieve near optimal generalization performance on test samples. Specifically it is shown that under suitable assumptions, Bayesian optimal robust estimator requires test-time adaptation, and such adaptation can lead to significant performance boost over standard adversarial training. Motivated by this observation, we propose a practically easy to implement method to improve the generalization performance of adversarially-trained networks via an additional self-supervised test-time adaptation step. We further employs a meta adversarial training method to find a good starting point for test-time adaptation, which incorporates the test-time adaptation procedure into the training phase and it strengthens the correlation between the pre-text tasks in self-supervised learning and the original classification task. Extensive empirical experiments on CIFAR10, STL10 and Tiny ImageNet using several different self-supervised tasks show that our method consistently improves the robust accuracy of standard adversarial training under different white-box and black-box attack strategies.

****

AutoShot: A Short Video Dataset and State-of-the-Art Shot Boundary Detection

Wentao Zhu,Xiufeng Xie,Wenxian Liu,Jincan Deng,Debing Zhang,Zhangyang Wang,Ji Liu

The short-form videos have explosive popularity and have dominated the new social media trends. Prevailing short-video platforms, e.g., TikTok, Instagram Reels, and YouTube Shorts, have changed the way we consume and create content. For video content creation and understanding, the shot boundary detection (SBD) is one of the most essential components in various scenarios. In this work, we release a new public Short video sHot bOundary deTection dataset, named SHOT, consisting of 853 complete short videos and 11,606 shot annotations, with 2,716 high quali ty shot boundary annotations in 200 test videos. Leveraging this new data wealth, we propose to optimize the model design for video SBD, by conducting neural architecture search in a search space encapsulating various advanced 3D ConvNets and Transformers. Our proposed approach, named AutoShot, achieves higher F1 scores than previous state-of-the-art approaches, e.g., outperforming TransNetV2 by 4.2%, when being derived and evaluated on our newly constructed SHOT dataset. Moreover, to validate the generalizability of the AutoShot architecture, we directly evaluate it on another three public datasets: ClipShots, BBC and RAI, and the F1 scores of AutoShot outperform previous state-of-the-art approaches by 1.1%, 0.9% and 1.2%, respectively. The SHOT dataset and code will be released.

****

Prototypical Calibration for Few-shot Learning of Language Models

Zhixiong Han,Yaru Hao,Li Dong,Yutao Sun,Furu Wei

In-context learning of GPT-like models has been recognized as fragile across different hand-crafted templates, and demonstration permutations. In this work, we propose prototypical calibration to adaptively learn a more robust decision boundary for zero- and few-shot classification, instead of greedy decoding. Concretely, our method first adopts Gaussian mixture distribution to estimate the prototypical clusters for all categories. Then we assign each cluster to the corresponding label by solving a weighted bipartite matching problem. Given an example, its prediction is calibrated by the likelihood of prototypical clusters. Experimental results show that prototypical calibration yields a substantial improvement on a diverse set of tasks. Extensive analysis across different scales also indicates that our method calibrates the decision boundary as expected, greatly improving the robustness of GPT to templates, permutations, and class imbalance.

****

NERDS: A General Framework to Train Camera Denoisers from Raw-RGB Noisy Image Pairs

Heewon Kim,Kyoung Mu Lee

We aim to train accurate denoising networks for smartphone/digital cameras from single noisy images. Downscaling is commonly used as a practical denoiser for low-resolution images. Based on this processing, we found that the pixel variance of the natural images is more robust to downscaling than the pixel variance of the camera noises. Intuitively, downscaling easily removes high-frequency noises than natural textures. To utilize this property, we can adopt noisy/clean image synthesis at low-resolution to train camera denoisers. On this basis, we propose a new solution pipeline -- NERDS that estimates camera noises and synthesizes noisy-clean image pairs from only noisy images. In particular, it first models the noise in raw-sensor images as a Poisson-Gaussian distribution, then estimates the noise parameters using the difference of pixel variances by downscaling. We formulate the noise estimation as a gradient-descent-based optimization problem through a reparametrization trick. We further introduce a new Image Signal Processor (ISP) estimation method that enables denoiser training in a human-readable RGB space by transforming the synthetic raw images to the style of a given RGB noisy image. The noise and ISP estimations utilize rich augmentation to synthesize image pairs for denoiser training. Experiments show that our NERDS can accurately train CNN-based denoisers (e.g., DnCNN, ResNet-style network) outperforming previous noise-synthesis-based and self-supervision-based denoisers in real datasets.

****

Communication-Efficient and Drift-Robust Federated Learning via Elastic Net

Seonhyeong Kim,jiheon woo,Daewon Seo,Yongjune Kim

Federated learning (FL) is a distributed method to train a global model over a set of local clients while keeping data localized, which reduces risks of privacy and security. FL framework faces important challenges including expensive communication cost and client drift problem. Leveraging the elastic net, we propose a communication-efficient and drift-robust FL framework to improve the communication efficiency and resolve the client drift problem. We repurpose two types of the elastic net regularizers (i.e., ℓ_1 and ℓ_2 penalties on the local model updates): (1) the ℓ_1-norm regularizer sparsifies the local updates t o enhance the communication efficiency and (2) the ℓ_2-norm regularizer attempts to resolve the client drift problem by limiting the impact of drifting local updates due to data heterogeneity. Our framework is general; hence, it can be integrated with prior FL techniques, e.g., FedAvg, FedProx, SCAFFOLD, and FedDyn. We show that our framework effectively resolves the communication cost problem and the client drift problem simultaneously.

****

Learning Hierarchical Protein Representations via Complete 3D Graph Networks

Limei Wang,Haoran Liu,Yi Liu,Jerry Kurtin,Shuiwang Ji

We consider representation learning for proteins with 3D structures. We build 3D graphs based on protein structures and develop graph networks to learn their representations. Depending on the levels of details that we wish to capture, protein representations can be computed at different levels, \emph{e.g.}, the amino acid, backbone, or all-atom levels. Importantly, there exist hierarchical relations among different levels. In this work, we propose to develop a novel hierarchical graph network, known as ProNet, to capture the relations. Our ProNet is very flexible and can be used to compute protein representations at different levels of granularity. By treating each amino acid as a node in graph modeling as well as harnessing the inherent hierarchies, our ProNet is more effective and efficient than existing methods. We also show that, given a base 3D graph network that is complete, our ProNet representations are also complete at all levels. Experimental results show that ProNet outperforms recent methods on most datasets. In addition, results indicate that different downstream tasks may require representations at different levels. Our code is publicly available as part of the DIG library (\url{https://github.com/divelab/DIG}).

****

Adversarial Attacks on Adversarial Bandits

Yuzhe Ma,Zhijin Zhou

We study a security threat to adversarial multi-armed bandit, in which an attacker perturbs the loss or reward signal to control the behavior of the victim bandit player. We show that the attacker is able to mislead any no-regret adversarial bandit algorithm into selecting a suboptimal target action in every but sublinear (T−o(T)) number of rounds, while incurring only sublinear (o(T)) cumulative attack cost. This result implies critical security concern in real-world bandit-based systems, e.g., in online recommendation, an attacker might be able to hijack the recommender system and promote a desired product. Our proposed attack algorithms require knowledge of only the regret rate, thus are agnostic to the concrete bandit algorithm employed by the victim player. We also derived a theoretical lower bound on the cumulative attack cost that any victim-agnostic attack algorithm must incur. The lower bound matches the upper bound achieved by our attack, which shows that our attack is asymptotically optimal.

****

Grounding High Dimensional Representation Similarity by Comparing Decodability and Network Performance

Lucas Hayne,Heejung Jung,Abhijit Suresh,R. McKell Carter

To understand and interpret neural networks, representation similarity metrics have been used to compare learned representations between and across networks. Recent experiments have compared these similarity metrics to find the best performing and the most robust metrics, noting that classic baselines perform surprisingly well. These experiments are mostly constrained to studying relatively low-dimensional representations because of the computational cost of prominent representation similarity metrics. We extend previous work to test representation similarity metrics on larger convolutional networks processing larger images. In order to make this work possible, we employ reformulated representation similarity metrics for use on very high-dimensional representations. Using these reformulated similarity metrics, we test how well each metric captures changes to representations induced by ablations in two popular convolutional networks. In order to ground the effects of changes to representations in function, we use linear decoding probes and network performance measures. These measures of function allow us to test how well similarity metrics capture changes in decodable information ve rsus changes in network performance. Linear decoding methods index available information in the representation, while network performance measures index the information used by the network. We show that all the tested representation similarity metrics significantly predict changes in network function and decodability. Within these metrics, on average, Procrustes and CKA outperform regularized CCA-based methods. All metrics predict decodability changes significantly better than they do network function. Procrustes and CKA do not outperform regularized CCA-based metrics for all network and functionality measure combinations. We add to the growing literature on representational similarity metrics to facilitate the improvement of current metrics for network interpretability.

****

Likelihood adjusted semidefinite programs for clustering heterogeneous data

Yubo Zhuang,Xiaohui Chen,Yun Yang

Clustering is a widely deployed unsupervised learning tool. Model-based clustering is a flexible framework to tackle data heterogeneity when the clusters have different shapes. Likelihood-based inference for mixture distributions often involves non-convex and high-dimensional objective functions, imposing difficult computational and statistical challenges. The classic expectation-maximization (EM) algorithm is a computationally thrifty iterative method that maximizes a surrogate function minorizing the log-likelihood of observed data in each iteration, which however suffers from bad local maxima even in the special case of the standard Gaussian mixture model with common isotropic covariance matrices. On the other hand, recent studies reveal that the unique global solution of a semidefinite programming (SDP) relaxed K-means achieves the information-theoretically sharp threshold for perfectly recovering the cluster labels under the standard Gaussian mixture model. In this paper, we extend the SDP approach to a general setting by integrating cluster labels as model parameters and propose an iterative likelihood adjusted SDP (iLA-SDP) method that directly maximizes the \emph{exact} observed likelihood in the presence of data heterogeneity. By lifting the cluster assignment to group-specific membership matrices, iLA-SDP avoids centroids estimation -- a key feature that allows exact recovery under well-separateness of centroids without being trapped by their adversarial configurations. Thus iLA-SDP is less sensitive than EM to initialization and more stable on high-dimensional data. Our numeric experiments demonstrate that iLA-SDP can achieve lower mis-clustering errors over several widely used clustering methods including K-means, SDP and EM algorithms.

****

RGI: robust GAN-inversion for mask-free image inpainting and unsupervised pixel-wise anomaly detection

Shancong Mou,Xiaoyi Gu,Meng Cao,Haoping Bai,Ping Huang,Jiulong Shan,Jianjun Shi

Generative adversarial networks (GANs), trained on a large-scale image dataset, can be a good approximator of the natural image manifold. GAN-inversion, using a pre-trained generator as a deep generative prior, is a promising tool for image restoration under corruptions. However, the performance of GAN-inversion can be limited by a lack of robustness to unknown gross corruptions, i.e., the restored image might easily deviate from the ground truth. In this paper, we propose a Robust GAN-inversion (RGI) method with a provable robustness guarantee to achieve image restoration under unknown \textit{gross} corruptions, where a small fraction of pixels are completely corrupted. Under mild assumptions, we show that the restored image and the identified corrupted region mask converge asymptotically to the ground truth. Moreover, we extend RGI to Relaxed-RGI (R-RGI) for generator fine-tuning to mitigate the gap between the GAN learned manifold and the true image manifold while avoiding trivial overfitting to the corrupted input image, which further improves the image restoration and corrupted region mask identification performance. The proposed RGI/R-RGI method unifies two important applications with state-of-the-art (SOTA) performance: (i) mask-free semantic inpainting, where the corruptions are unknown missing regions, the restored background can be used to restore the missing content. (ii) unsupervised pixel-wise anomaly detection, where the corruptions are unknown anomalous regions, the retrieved mask can be used as the anomalous region’s segmentation mask.

****

Coverage-centric Coreset Selection for High Pruning Rates

Haizhong Zheng,Rui Liu,Fan Lai,Atul Prakash

One-shot coreset selection aims to select a representative subset of the training data, given a pruning rate, that can later be used to train future models while retaining high accuracy. State-of-the-art coreset selection methods pick the highest importance examples based on an importance metric and are found to perform well at low pruning rates. However, at high pruning rates, they suffer from a catastrophic accuracy drop, performing worse than even random sampling. This paper explores the reasons behind this accuracy drop both theoretically and empirically. We first propose a novel metric to measure the coverage of a dataset on a specific distribution by extending the classical geometric set cover problem to a distribution cover problem. This metric helps explain why coresets selected by SOTA methods at high pruning rates perform poorly compared to random sampling because of worse data coverage. We then propose a novel one-shot coreset selection method, Coverage-centric Coreset Selection (CCS), that jointly considers overall data coverage upon a distribution as well as the importance of each example. We evaluate CCS on five datasets and show that, at high pruning rates (e.g., 90%), it achieves significantly better accuracy than previous SOTA methods (e.g., at least 19.56% higher on CIFAR10) as well as random selection (e.g., 7.04% higher on CIFAR10) and comparable accuracy at low pruning rates. We make our code publicly available at https://github.com/haizhongzheng/Coverage-centric-coreset-selection.

****

AIA: learn to design greedy algorithm for NP-complete problems using neural networks

Zhenxin Ding,Jingyan Sui,Ruizhi Liu,Shizhe Ding,Liming Xu,Zihao Huang,Chao Wang,Haicang Zhang,Shiwei Sun,Chungong Yu,Dongbo Bu

Algorithm design is an art that heavily requires intuition and expertise of the human designers as well as insights into the problems under consideration. In particular, the design of greedy-selection rules, the core of greedy algorithms, is usually a great challenge to designer: it is relatively easy to understand a greedy algorithm while it is always difficult to find out an effective greedy-selection rule. In the study, we present an approach, called AIA, to learn algorithm design with the aid of neural networks. We consider the minimum weighted set cover (SC) problem, one of the NP-hard problems, as an representative example. Initially, we formulate a given weighted SC problem as an 0-1 integer linear program (ILP): each variable x_i has two options, i.e., $x_i=0$, which denotes abandon of the set s_i, and $x_i = 1$, which denotes selection of s_i. Each option of a variable leads to a sub-problem with respect to the original ILP problem. Next, we design a generic search framework to find the optimal solution to the ILP problem. At each search step, the value of a variable is determined with the aid of neural networks. The key of our neural network is the loss function: the original ILP problem and the sub-problems generated by assigning a variable x_i should satisfy the Bellman-Ford equation, and the dissatisfication of the Bellman-Ford equation is evaluated and used as loss function of our neural network. The trained neural network is used as greedy-selection rule. Experimental results on representative instances suggest that using the NN-based greedy selection rule, we can successfully find the optimal solutions. More importantly, the NN-based greedy-selection rule outperform the outstanding Chavatal greedy algorithm, which was designed by human expert. The basic idea of our approach can be readily extended without significant modification to design greedy algorithm for other NP-hard problems.

****

Self-Adaptive Perturbation Radii for Adversarial Training

Huimin Wu,Zhang Chenkang,Bin Gu

Adversarial training has been shown to be the most popular and effective technique to protect models from imperceptible adversarial samples. Despite its success, it also accompanies the significant performance degeneration to clean data. To achieve a good performance on both clean and adversarial samples, the main eff ort is searching for an adaptive perturbation radius for each training sample, which essentially suffers from a conflict between exact searching and computational overhead. To address this conflict, in this paper, firstly we show the superiority of adaptive perturbation radii intuitively and theoretically regarding the accuracy and robustness respectively. Then we propose our novel self-adaptive adjustment framework for perturbation radii without tedious searching. We also discuss this framework on both deep neural networks (DNNs) and kernel support vector machines (SVMs). Finally, extensive experimental results show that our framework can improve not only natural generalization performance but also adversarial robustness. It is also competitive with existing searching strategies in terms of running time.

****

GCINT: Dynamic Quantization Algorithm for Training Graph Convolution Neural Networks Using Only Integers

Qizhe Wu,Letian Zhao,Huawen Liang,Xiaotian Wang,LinFeng Tao,Teng Tian,Tingxin Wang,Zerong He,Wei Wu,Xi Jin

Quantization approaches can minimize storage costs while decreasing the computational complexity of a model, although there is minimal study in the GNN field on quantization networks. We studied the four primary reasons why existing quantization approaches cannot be employed extensively with GNNs: (1)Quantifying the distinctions between data sources; (2)Quantifying the distinctions between data streams; (3)Quantifying the distinctions between concentrations; (4)QAT’s Limitations. Based on this, we propose GCINT, which is an efficient quantization framework prepared for GNN training. The entire forward, backward, optimizer, and loss functions are calculated using integer data. We achieved a training acceleration ratio of nearly 10× compared to FP32 Cuda Core in RTX 2080TI INT8 Tensor Core. Our quantization is independent of the dataset and weight distribution, and more than 2,000 randomized trials have been undertaken on the 8 popular GNN benchmark datasets, with all achieving errors within 1% of the FP32.

****

ILA-DA: Improving Transferability of Intermediate Level Attack with Data Augmentation

Chiu Wai Yan,Tsz-Him Cheung,Dit-Yan Yeung

Adversarial attack aims to generate deceptive inputs to fool a machine learning model. In deep learning, an adversarial input created for a specific neural network can also trick other neural networks. This intriguing property is known as black-box transferability of adversarial examples. To improve black-box transferability, a previously proposed method called Intermediate Level Attack (ILA) fine-tunes an adversarial example by maximizing its perturbation on an intermediate layer of the source model. Meanwhile, it has been shown that simple image transformations can also enhance attack transferability. Based on these two observations, we propose ILA-DA, which employs three novel augmentation techniques to enhance ILA. Specifically, we propose (1) an automated way to apply effective image transformations, (2) an efficient reverse adversarial update technique, and (3) an attack interpolation method to create more transferable adversarial examples. Shown by extensive experiments, ILA-DA greatly outperforms ILA and other state-of-the-art attacks by a large margin. On ImageNet, we attain an average attack success rate of 84.5%, which is 19.5% better than ILA and 4.7% better than the previous state-of-the-art across nine undefended models. For defended models, ILA-DA also leads existing attacks and provides further gains when incorporated into more advanced attack methods.

****

Contrastive Alignment of Vision to Language Through Parameter-Efficient Transfer Learning

Zaid Khan,Yun Fu

Contrastive vision-language models (e.g. CLIP) are typically created by updating all the parameters of a vision model and language model through contrastive training. Can such models be created by a small number of parameter updates to an already-trained language model and vision model? The literature describes techniques that can create vision-language models by updating a small number of paramet

ers in a language model, but these require already aligned visual representations and are non-contrastive, hence unusable for latency-sensitive applications such as neural search. We explore the feasibility and benefits of parameter-efficient contrastive vision-language alignment through transfer learning: creating a model such as CLIP by minimally updating an already-trained vision and language model. We find that a minimal set of parameter updates ($<$7\%) can achieve the same performance as full-model training, and updating specific components ($<$1\% of parameters) can match 75\% of full-model training. We describe a series of experiments: we show that existing knowledge is conserved more strongly in parameter-efficient training and that parameter-efficient scaling scales with model and dataset size. Where paired-image text data is scarce but strong multilingual language models exist (e.g. low resource languages), parameter-efficient training is even preferable to full-model training. Given a fixed compute budget, parameter-efficient training allows training larger models on the same hardware, achieving equivalent performance in less time. Parameter-efficient training hence constitutes an energy-efficient and effective training strategy for contrastive vision-language models that may be preferable to the full-model training paradigm for common use cases.
Code and weights at https://github.com/codezakh/LilT.

****

BEEF: Bi-Compatible Class-Incremental Learning via Energy-Based Expansion and Fusion

Fu-Yun Wang,Da-Wei Zhou,Liu Liu,Han-Jia Ye,Yatao Bian,De-Chuan Zhan,Peilin Zhao

Neural networks suffer from catastrophic forgetting when sequentially learning tasks phase-by-phase, making them inapplicable in dynamically updated systems. Class-incremental learning (CIL) aims to enable neural networks to learn different categories at multi-stages. Recently, dynamic-structure-based CIL methods achieve remarkable performance. However, these methods train all modules in a coupled manner and do not consider possible conflicts among modules, resulting in spoilage of eventual predictions. In this work, we propose a unifying energy-based theory and framework called Bi-Compatible Energy-Based Expansion and Fusion (BEEF) to analyze and achieve the goal of CIL. We demonstrate the possibility of training independent modules in a decoupled manner while achieving bi-directional compatibility among modules through two additionally allocated prototypes, and then integrating them into a unifying classifier with minimal cost. Furthermore, BEEF extends the exemplar-set to a more challenging setting, where exemplars are randomly selected and imbalanced, and maintains its performance when prior methods fail dramatically.
Extensive experiments on three widely used benchmarks: CIFAR-100, ImageNet-100, and ImageNet-1000 demonstrate that BEEF achieves state-of-the-art performance in both the ordinary and challenging CIL settings. The Code is available at https://github.com/G-U-N/ICLR23-BEEF.

****

Out-of-distribution Representation Learning for Time Series Classification

Wang Lu,Jindong Wang,Xinwei Sun,Yiqiang Chen,Xing Xie

Time series classification is an important problem in the real world. Due to its non-stationary property that the distribution changes over time, it remains challenging to build models for generalization to unseen distributions. In this paper, we propose to view time series classification from the distribution perspective. We argue that the temporal complexity of a time series dataset could attribute to unknown latent distributions that need characterize. To this end, we propose DIVERSIFY for out-of-distribution (OOD) representation learning on dynamic distributions of times series. DIVERSIFY takes an iterative process: it first obtains the ‘worst-case’ latent distribution scenario via adversarial training, then reduces the gap between these latent distributions. We then show that such an algorithm is theoretically supported. Extensive experiments are conducted on seven datasets with different OOD settings across gesture recognition, speech commands recognition, wearable stress and affect detection, and sensor-based human activity recognition. Qualitative and quantitative results demonstrate that DIVERSIFY significantly outperforms other baselines and effectively characterizes the latent distributions. Code is available at https://github.com/microsoft/robustlearn.

****

A Closer Look at the Calibration of Differentially Private Learners
Hanlin Zhang,Xuechen Li,Prithviraj Sen,Salim Roukos,Tatsunori Hashimoto

We systematically study the calibration of classifiers trained with differentially private stochastic gradient descent (DP-SGD) and observe miscalibration across a wide range of vision and language tasks. Our analysis identifies per-example gradient clipping in DP-SGD as a major cause of miscalibration, and we show that existing approaches for improving calibration with differential privacy only provide marginal improvements in calibration error while occasionally causing large degradations in accuracy. As a solution, we show that differentially private variants of post-processing calibration methods such as temperature scaling and Platt scaling are surprisingly effective and have negligible utility cost to the overall model. Across 7 tasks, temperature scaling and Platt scaling with DP-SGD result in an average 3.1-fold reduction in the in-domain expected calibration error and only incur at most a minor percent drop in accuracy.

****

Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation
YiFan Zhang,Hanlin Zhang,Zachary Chase Lipton,Li Erran Li,Eric Xing

Previous works on Treatment Effect Estimation (TEE) are not in widespread use because they are predominantly theoretical, where strong parametric assumptions are made but untractable for practical application. Recent works use Multilayer Perceptron (MLP) for modeling casual relationships, however, MLPs lag far behind recent advances in ML methodology, which limits their applicability and generalizability. To extend beyond the single domain formulation and towards more realistic learning scenarios, we explore model design spaces beyond MLPs, i.e., transformer backbones, which provide flexibility where attention layers govern interactions among treatments and covariates to exploit structural similarities of potential outcomes for confounding control. Through careful model design, Transformers as Treatment Effect Estimators (TransTEE) is proposed. We show empirically that TransTEE can: (1) serve as a general-purpose treatment effect estimator which significantly outperforms competitive baselines on a variety of challenging TEE problems (e.g., discrete, continuous, structured, or dosage-associated treatments.) and is applicable to both when covariates are tabular and when they consist of structural data (e.g., texts, graphs); (2) yield multiple advantages: compatibility with propensity score modeling, parameter efficiency, robustness to continuous treatment value distribution shifts, explainable in covariate adjustment, and real-world utility in auditing pre-trained language models.

****

Few-Shot Learning with Representative Global Prototype
Yukun Liu,Daming Shi,Hexiu Lin

Few-shot learning is often challenged by low generalization performance due to the assumption that the data distribution of novel classes and base classes is similar while the model is trained only on the base classes. To mitigate the above issues, we propose a few-shot learning with representative global prototype method. Specifically, to enhance the generalization to novel classes, we propose a method to jointly train the base classes and the novel classes, using selected representative and non-representative samples to optimize representative global prototypes, respectively. Additionally, a method that organically combines the sample of base classes conditional on semantic embedding to generate new samples of novel classes with the original data is proposed to enhance the data of novel classes. Results show that this training method improves the model's ability to describe novel classes, improving the classification performance for a few shots. Intensive experiments have been conducted on two popular benchmark datasets, and the experimental results show that this method significantly improves the classification ability of few-shot learning tasks and achieves state-of-the-art performance.

****

Feature-Driven Talking Face Generation with StyleGAN2

Tao Zhang,Kai Tang,Weiwu Zhang,Kazushige Ouchi

In this work, we wish to use a face image that generate a more natural and real face talking animation video. This is not an easy task because face appearance variation and semantics of speech are coupled together when tacking face have a micro movement. Audio features sometimes contain information about expressions, but they are not accurate enough. So a single audio feature cannot fully represent the movement of the face. For the above reason, we want to use different features to generate talking faces. The StyleGan series show good performance in the direction of image processing, and can perform the style migration task of portraits very well at the same time. We find that StyleGan can be used as a talking face generator. At the same time, we also encode and extract non-identity features and non-lip features, and try to find the subtle relationship between the features and the talking face. We also use the evaluation and ablation study to measure the quality of the generated videos and examine whether our approach is effective and feasible.

****

Schema Inference for Interpretable Image Classification

Haofei Zhang,Mengqi Xue,Xiaokang Liu,Kaixuan Chen,Jie Song,Mingli Song

In this paper, we study a novel inference paradigm, termed as schema inference, that learns to deductively infer the explainable predictions by rebuilding the prior deep neural network (DNN) forwarding scheme, guided by the prevalent philosophical cognitive concept of schema. We strive to reformulate the conventional model inference pipeline into a graph matching policy that associates the extracted visual concepts of an image with the pre-computed scene impression, by analogy with human reasoning mechanism via impression matching. To this end, we devise an elaborated architecture, termed as SchemaNet, as a dedicated instantiation of the proposed schema inference concept, that models both the visual semantics of input instances and the learned abstract imaginations of target categories as topological relational graphs. Meanwhile, to capture and leverage the compositional contributions of visual semantics in a global view, we also introduce a universal Feat2Graph scheme in SchemaNet to establish the relational graphs that contain abundant interaction information. Both the theoretical analysis and the experimental results on several benchmarks demonstrate that the proposed schema inference achieves encouraging performance and meanwhile yields a clear picture of the deductive process leading to the predictions. Our code is available at https://github.com/zhfeing/SchemaNet-PyTorch.

****

Supernet Training for Federated Image Classification Under System Heterogeneity

Taehyeon Kim,Se-Young Yun

Efficient deployment of deep neural networks across many devices and resource constraints, particularly on edge devices, is one of the most challenging problems in the presence of data-privacy preservation issues. Conventional approaches have evolved to either improve a single global model while keeping each local heterogeneous training data decentralized (i.e. data heterogeneity; Federated Learning (FL)) or to train an overarching network that supports diverse architectural settings to address heterogeneous systems equipped with different computational capabilities (i.e. system heterogeneity; Neural Architecture Search). However, few studies have considered both directions simultaneously. This paper proposes the federation of supernet training (FedSup) framework to consider both scenarios simultaneously, i.e., where clients send and receive a supernet that contains all possible architectures sampled from itself. The approach is inspired by observing that averaging parameters during model aggregation for FL is similar to weight-sharing in supernet training. Thus, the proposed FedSup framework combines a weight-sharing approach widely used for training single shot models with FL averaging (FedAvg). Furthermore, we develop an efficient algorithm (E-FedSup) by sending the sub-model to clients on the broadcast stage to reduce communication costs and training overhead, including several strategies to enhance supernet training in the FL environment. We verify the proposed approach with extensive empirical evaluations. The resulting framework also ensures data and model heterogeneity robustness on several standard benchmarks.

****

CircuitNet: A Generic Neural Network to Realize Universal Circuit Motif Modeling
Yansen Wang,XINYANG JIANG,Kan Ren,Caihua Shan,Xufang Luo,Kaitao Song,Dongsheng Li

The successes of artificial neural networks (ANNs) are largely attributed to mimicking the human brain structures. Recent advances in neuroscience revealed that neurons interact with each other through various kinds of connectivity patterns to process information, in which the common connectivity patterns are also called circuit motifs. However, many existing ANNs can only model one or two circuit motifs in their architectures, so that their performance may drastically vary among different types of machine learning tasks.

In this paper, we propose a new type of neural network inspired by the architectures of neuronal circuits, namely Circuit Neural Network (CircuitNet). In CircuitNet, a group of densely connected neurons, namely circuit motif unit (CMU), form the basic unit of the network, which is capable of modeling universal circuit motifs by adjusting the weights within the CMUs. Compared with traditional feed-forward networks, CircuitNet has the ability to model more types of neuron connections such as feed-back and lateral motifs.

Inspired by the locally dense and globally sparse structure of the human brain, several iterations of signal transmission among different CMUs are achieved by sparse connections through the input ports and output ports of different CMUs.

Experiments have demonstrated that CircuitNet can outperform popular neural network architectures in function approximation, reinforcement learning, image classification, and time series forecasting tasks.

****

Your Contrastive Learning Is Secretly Doing Stochastic Neighbor Embedding
Tianyang Hu,Zhili LIU,Fengwei Zhou,Wenjia Wang,Weiran Huang

Contrastive learning, especially self-supervised contrastive learning (SSCL), has achieved great success in extracting powerful features from unlabeled data. In this work, we contribute to the theoretical understanding of SSCL and uncover its connection to the classic data visualization method, stochastic neighbor embedding (SNE), whose goal is to preserve pairwise distances. From the perspective of preserving neighboring information, SSCL can be viewed as a special case of SNE with the input space pairwise similarities specified by data augmentation. The established correspondence facilitates deeper theoretical understanding of learned features of SSCL, as well as methodological guidelines for practical improvement. Specifically, through the lens of SNE, we provide novel analysis on domain-agnostic augmentations, implicit bias and robustness of learned features. To illustrate the practical advantage, we demonstrate that the modifications from SNE to t-SNE can also be adopted in the SSCL setting, achieving significant improvement in both in-distribution and out-of-distribution generalization.

****

Covariance-Robust Minimax Probability Machines for Algorithmic Recourse
Ngoc Bui,Duy Nguyen,Kim-Cuc Nguyen,Man-Chung Yue,Viet Anh Nguyen

Algorithmic recourse is rising as a prominent technique to promote the explainability and transparency of the predictive model in ethical machine learning. Existing approaches to algorithmic recourse often assume an invariant predictive model; however, this model, in reality, is usually updated temporally upon the input of new data. Thus, a recourse that is valid respective to the present model may become invalid for the future model. To resolve this issue, we propose a pipeline to generate a model-agnostic recourse that is robust to model shifts. Our pipeline first estimates a linear surrogate of the nonlinear (black-box) model using covariance-robust minimax probability machines (MPM); then, the recourse is generated with respect to this robust linear surrogate. We show that the covariance-robust MPM recovers popular regularization schemes, including l_2-regularization and class-reweighting. We also show that our covariance-robust MPM pushes the decision boundary in an intuitive manner, which facilitates an interpretable generation of a robust recourse. The numerical results demonstrate the usefulness and robustness of our pipeline.

****

Harnessing Mixed Offline Reinforcement Learning Datasets via Trajectory Weighting

Zhang-Wei Hong,Pulkit Agrawal,Remi Tachet des Combes,Romain Laroche

Most offline reinforcement learning (RL) algorithms return a target policy maximizing a trade-off between (1) the expected performance gain over the behavior policy that collected the dataset, and (2) the risk stemming from the out-of-distribution-ness of the induced state-action occupancy. It follows that the performance of the target policy is strongly related to the performance of the behavior policy and, thus, the trajectory return distribution of the dataset. We show that in mixed datasets consisting of mostly low-return trajectories and minor high-return trajectories, state-of-the-art offline RL algorithms are overly restrained by low-return trajectories and fail to exploit high-performing trajectories to the fullest. To overcome this issue, we show that, in deterministic MDPs with stochastic initial states, the dataset sampling can be re-weighted to induce an artificial dataset whose behavior policy has a higher return. This re-weighted sampling strategy may be combined with any offline RL algorithm. We further analyze that the opportunity for performance improvement over the behavior policy correlates with the positive-sided variance of the returns of the trajectories in the dataset. We empirically show that while CQL, IQL, and TD3+BC achieve only a part of this potential policy improvement, these same algorithms combined with our reweighted sampling strategy fully exploit the dataset. Furthermore, we empirically demonstrate that, despite its theoretical limitation, the approach may still be efficient in stochastic environments.

****

Self-Consistency Improves Chain of Thought Reasoning in Language Models

Xuezhi Wang,Jason Wei,Dale Schuurmans,Quoc V Le,Ed H. Chi,Sharan Narang,Aakanksha Chowdhery,Denny Zhou

Chain-of-thought prompting combined with pretrained large language models has achieved encouraging results on complex reasoning tasks. In this paper, we propose a new decoding strategy, self-consistency, to replace the naive greedy decoding used in chain-of-thought prompting. It first samples a diverse set of reasoning paths instead of only taking the greedy one, and then selects the most consistent answer by marginalizing out all possible reasoning paths. Self-consistency leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer. Our extensive empirical evaluation shows that self-consistency boosts the performance of chain-of-thought prompting with a striking margin on a range of popular arithmetic and commonsense reasoning benchmarks, including GSM8K (+17.9%), SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%) and ARC-challenge (+3.9%).

****

Ensuring DNN Solution Feasibility for Optimization Problems with Linear Constraints

Tianyu Zhao,Xiang Pan,Minghua Chen,Steven Low

We propose preventive learning as the first framework to guarantee Deep Neural Network (DNN) solution feasibility for optimization problems with linear constraints without post-processing, upon satisfying a mild condition on constraint calibration. Without loss of generality, we focus on problems with only inequality constraints. We systematically calibrate the inequality constraints used in training, thereby anticipating DNN prediction errors and ensuring the obtained solutions remain feasible. We characterize the calibration rate and a critical DNN size, based on which we can directly construct a DNN with provable solution feasibility guarantee. We further propose an Adversarial-Sample Aware training algorithm to improve its optimality performance. We apply the framework to develop DeepOPF+ for solving essential DC optimal power flow problems in grid operation. Simulation results over IEEE test cases show that it outperforms existing strong DNN baselines in ensuring 100\% feasibility and attaining consistent optimality loss (<0.19%) and speedup (up to x228) in both light-load and heavy-load regimes, as compared to a state-of-the-art solver. We also apply our framework to a non-convex problem and show its performance advantage over existing schemes.

****

EM-Network: Learning Better Latent Variable for Sequence-to-Sequence Models

Ji Won Yoon,SungHwan Ahn,Hyeonseung Lee,Minchan Kim,SeokMin Kim,Nam Soo Kim

In a sequence-to-sequence (seq2seq) framework, the use of an unobserved latent variable, such as latent alignment and representation, is important to address the mismatch problem between the source input and target output sequences. Existing seq2seq literature typically learns the latent space by only consuming the source input, which might produce a sub-optimal latent variable for predicting the target. Extending an expectation-maximization (EM)-like algorithm, we introduce EM-Network that can yield the promising latent variable by leveraging the target sequence as the model's additional training input. The target input is used as guidance to provide the target-side context and reduce the candidates of the latent variable. The proposed framework is trained in a new self-distillation setup, allowing the original sequence model to benefit from the latent variable of the EM-Network. Specifically, the EM-Network's prediction serves as a soft label for training the inner sequence model, which only takes the source as input. We theoretically show that our training objective can be a lower bound for the log-likelihood of the sequence model and is justified from the EM perspective. We conduct comprehensive experiments on two sequence learning tasks: speech recognition and machine translation. Experimental results demonstrate that the EM-Network significantly advances the current state-of-the-art self-supervised learning approaches. It improves over the best prior work on speech recognition and establishes state-of-the-art performance on WMT'14 and IWSLT'14 datasets. Moreover, the proposed method even achieves considerable performance improvement for fully supervised learning.

****

AutoFHE: Automated Adaption of CNNs for Efficient Evaluation over FHE

Wei Ao,Vishnu Boddeti

Secure inference of deep convolutional neural networks (CNNs) was recently demonstrated under the RNS-CKKS fully homomorphic encryption (FHE) scheme. The state-of-the-art solution uses a high-order composite polynomial to approximate non-arithmetic ReLUs and refreshes zero-level ciphertext through bootstrapping. However, this solution suffers from prohibitively high latency, both due to the number of levels consumed by the polynomials (47%) and the inference time consumed by bootstrapping operations (70%). Furthermore, it requires a hand-crafted architecture for homomorphically evaluating CNNs by placing a bootstrapping operation after every Conv-BN layer. To accelerate CNNs on FHE and automatically design a homomorphic evaluation architecture, we propose AutoFHE: Automated adaption of CNNs for evaluation over FHE. AutoFHE exploits the varying sensitivity of approximate activations across different layers in a network and jointly evolves polynomial activations (EvoReLUs) and searches for placement of bootstrapping operations for evaluation under RNS-CKKS. The salient features of AutoFHE include: i) a multi-objective co-evolutionary (MOCoEv) search algorithm to maximize validation accuracy and minimize the number of bootstrapping operations, ii) a gradient-free search algorithm, R-CCDE, to optimize EvoReLU coefficients, and iii) polynomial-aware training (PAT) to fine-tune polynomial-only CNNs for one epoch to adapt trainable weights to EvoReLUs. We demonstrate the efficacy of AutoFHE through the evaluation of ResNets on CIFAR-10 and CIFAR-100 under RNS-CKKS. Experimental results on CIFAR-10 indicate that in comparison to the state-of-the-art solution, AutoFHE reduces inference time (50 images on 50 threads) by 1,000 seconds and amortized inference time (per image) by 28% and 17% for ResNet-20 and ResNet-32, respectively.

****

REPRESENTATIVE PROTOTYPE WITH CONSTRASTIVE LEARNING FOR SEMI-SUPENVISED FEW-SHOT CLASSIFICATION

Hexiu Lin,Daming Shi,Yukun Liu

Few-shot learning aims to learn novel classes in the dataset with few samples per class, which is a very challenging task. To mitigate this issue, the prior work obtain representative prototypes with semantic embeddin based on prototypical networks. While the above methods do not meet the requirement of few-shot learning, which requires abundant labeled samples. Therefore, We propose a new model f ramework to get representative prototypes with semi-supervised learning. Specifically, we introduces the dataset containing unlabeled samples to assist training the model. More importantly, to fully utilize these unlabeled samples, we adopt conditional variational autoencoder to construct more representative prototypes. Simultaneously, we develop novel contrastive loss to improve the model generalization ability. We evaluate our method on miniImageNet and tieredImageNet benchmarks for both 1-shot and 5-shot settings and achieve better performance over the state-of-the-art semi■supervised few-shot method.

****

Data-efficient Supervised Learning is Powerful for Neural Combinatorial Optimization

Shunyu Yao,Xi Lin,Zhenkun Wang,Qingfu Zhang

Neural combinatorial optimization (NCO) is a promising learning-based approach to solve difficult combinatorial optimization problems. However, how to efficiently train a powerful NCO solver remains challenging. The widely-used reinforcement learning method suffers from sparse rewards and low data efficiency, while the supervised learning approach requires a large number of high-quality solutions. In this work, we develop efficient methods to extract sufficient supervised information from limited labeled data, which can significantly overcome the main shortcoming of supervised learning. To be specific, we propose a set of efficient data augmentation methods and a novel bidirectional loss to better leverage the equivalent properties of problem instances, which finally lead to a promising supervised learning approach. The thorough experimental studies demonstrate our proposed method can achieve state-of-the-art performance on the traveling salesman problem (TSP) only with a small set of 50,000 labeled instances, while it also enjoys better generalization performance. We believe this somewhat surprising finding could lead to valuable rethinking on the value of efficient supervised learning for NCO.

****

Temporally-Weighted Spike Encoding for Event-based Object Detection and Classification

Nikolaus Salvatore,Justin Fletcher

Event-based cameras exhibit high dynamic range and temporal precision that could make them ideal for detecting objects with high speeds and low relative luminance. These properties have made event-based cameras especially interesting for use in space domain awareness tasks, such as detecting dim, artificial satellites with high brightness backgrounds using ground-based optical sensors; however, the asynchronous nature of event-based data presents new challenges to performing objection detection. While spiking neural networks (SNNs) have been shown to naturally complement the asynchronous and binary properties of event-based data, they also present a number of challenges in their training, such as the spike vanishing problem and the large number of timesteps required for maximizing classification and detection accuracy. Furthermore, the extremely high sampling rate of event-based sensors and the density of noisy space-based data collections can results in excessively large event streams within a short window of recording. We present a temporally-weighted spike encoding that greatly reduces the number of spikes derived from an event-based data stream, enabling the training of larger SNNs with fewer timesteps for maximal accuracy. We propose using this spike encoding with a variant of convolutional SNN trained utilizing surrogate spiking neuron gradients with backpropagation-through-time (BPTT) for both classification and object detection tasks with an emphasis on space-domain awareness. To demonstrate the efficacy of our encoding and SNN approach, we present competitive classification accuracies on benchmark datasets N-MNIST (99.7%), DVS-CIFAR10 (74.0%), and N-Caltech101 (72.8%), as well as state-of-the-art object detection performance on event-based, satellite collections.

****

Spiking Convolutional Neural Networks for Text Classification

Changze Lv,Jianhan Xu,Xiaoqing Zheng

Spiking neural networks (SNNs) offer a promising pathway to implement deep neural networks (DNNs) in a more energy-efficient manner since their neurons are spar sely activated and inferences are event-driven. However, there have been very few works that have demonstrated the efficacy of SNNs in language tasks partially because it is non-trivial to represent words in the forms of spikes and to deal with variable-length texts by SNNs. This work presents a "conversion + fine-tuning'' two-step method for training SNN for text classification and proposes a simple but effective way to encode pre-trained word embeddings as spike trains. We show empirically that after further fine-tuning with surrogate gradients, the converted SNNs achieve comparable results to their DNN counterparts across multiple datasets for Both English and Chinese. We also demonstrate that such SNNs are more robust against adversarial attacks than DNNs.

****

Personalized Federated Learning with Feature Alignment and Classifier Collaboration

Jian Xu,Xinyi Tong,Shao-Lun Huang

Data heterogeneity is one of the most challenging issues in federated learning, which motivates a variety of approaches to learn personalized models for participating clients. One such approach in deep neural networks based tasks is employing a shared feature representation and learning a customized classifier head for each client. However, previous works do not utilize the global knowledge during local representation learning and also neglect the fine-grained collaboration between local classifier heads, which limits the model generalization ability. In this work, we conduct explicit local-global feature alignment by leveraging global semantic knowledge for learning a better representation. Moreover, we quantify the benefit of classifier combination for each client as a function of the combining weights and derive an optimization problem for estimating optimal weights. Finally, extensive evaluation results on benchmark datasets with various heterogeneous data scenarios demonstrate the effectiveness of our proposed method.

****

Distributionally Robust Recourse Action

Duy Nguyen,Ngoc Bui,Viet Anh Nguyen

A recourse action aims to explain a particular algorithmic decision by showing one specific way in which the instance could be modified to receive an alternate outcome. Existing recourse generation methods often assume that the machine learning model does not change over time. However, this assumption does not always hold in practice because of data distribution shifts, and in this case, the recourse action may become invalid. To redress this shortcoming, we propose the Distributionally Robust Recourse Action (DiRRAc) framework, which generates a recourse action that has high probability of being valid under a mixture of model shifts. We first formulate the robustified recourse setup as a min-max optimization problem, where the max problem is specified by Gelbrich distance over an ambiguity set around the distribution of model parameters. Then we suggest a projected gradient descent algorithm to find a robust recourse according to the min-max objective. We also show that our DiRRAc framework can be extended to hedge against the misspecification of the mixture weights. Numerical experiments with both synthetic and three real-world datasets demonstrate the benefits of our proposed framework over the state-of-the-art recourse methods, which generate robust recourses.

****

Rethinking the Structure of Stochastic Gradients: Empirical and Statistical Evidence

Zeke Xie,Qian-Yuan Tang,Zheng He,Mingming Sun,Ping Li

It is well known that stochastic gradients significantly improve both optimization and generalization of deep neural networks (DNNs). Some works attempted to explain the success of stochastic optimization for deep learning by the arguably heavy-tail properties of gradient noise, while other works presented theoretical and empirical evidence against the heavy-tail hypothesis on gradient noise. Unfortunately, formal statistical tests for analyzing the structure and heavy tails of stochastic gradients in deep learning are still under-explored. In this paper, we mainly make two contributions. First, we conduct formal statistical tests o n the distribution of stochastic gradients and gradient noise across both parameters and iterations. Our statistical tests reveal that dimension-wise gradients usually exhibit power-law heavy tails, while iteration-wise gradients and stochastic gradient noise caused by minibatch training usually do not exhibit power-law heavy tails. Second, we further discover that the covariance spectra of stochastic gradients have the power-law structures in deep learning. While previous papers believed that the anisotropic structure of stochastic gradients matters to deep learning, they did not expect the gradient covariance can have such an elegant mathematical structure. Our work challenges the existing belief and provides novel insights on the structure of stochastic gradients. The novel structure of stochastic gradients may help understand the success of stochastic optimization for deep learning.

****

Representing Multi-view Time-series Graph Structures for Multivariate Long-term Time-series Forecasting

Wzh Rslh,Jin Fan,Huifeng Wu,Danfeng Sun

Multivariate long-term time-series forecasting task is a very challenging task in real-world application areas, such as electricity consumption and influenza-like illness forecasting. At present, researchers are focusing on designing robust and effective models, and have achieved good results. However, there are several issues with existing models that need to be overcome to ensure they provide optimal performance. First, the lack of a relationship structure between multivariate variables needs to be addressed. Second, most models only have a weak ability to capture local dynamic changes across the entire long-term time-series. And, third, the current models suffer from high computational complexity and unsatisfactory accuracy. To address these issues, we propose a novel method called Multi-view Time-series Graph Structure Representation (MTGSR) for multivariate long-term time-series forecasting tasks. MTGSR uses graph convolutional networks (GCNs) to construct topological relationships in the multivariate long-term time-series from three different perspectives: time, dimension, and crossing segments. Variation trends in the different dimensions of the multivariate long-term time-series are extracted through a difference operation so as to construct a topological map that reflects the correlations between the different dimensions. Then, to capture the dynamically changing characteristics of the fluctuation correlations between adjacent local sequences, MTGSR constructs a cross graph by calculating the correlation coefficients between adjacent local sequences. Extensive experiments on five different datasets show that MTGSR reduces errors by 20.41% over the state-of-the-art while maintaining linear complexity. Additionally, memory use is decreased by 66.52% and running time is reduced by 78.09%.

****

Improving Language Model Pretraining with Text Structure Information

Yi-Siang Wang,Ryohei Sasano,Koichi Takeda

Inter-sentence pretraining tasks learn from sentence relationships and facilitate high-level language understanding that cannot be directly learned in word-level pretraining tasks. However, we have found experimentally that existing inter-sentence methods for general-purpose language pretraining improve performance only at a relatively small scale but not at larger scales. For an alternative, we propose Text Structure Prediction (TSP), a more sophisticated inter-sentence task that uses text structure to provide more abundant self-supervised learning signals to pretraining models at larger scales. TSP classifies sentence pairs over six designed text structure relationships and it can be seen as an implicit form of learning high-level language understanding by identifying key concepts and relationships in texts. Experiments show that TSP provides improved performance on language understanding tasks for models at various scales. Our approach thus serves as an initial attempt to demonstrate that the exploitation of text structure can facilitate language understanding.

****

Chasing Better Deep Image Priors Between Over- and Under-parameterization

Qiming Wu,Xiaohan Chen,Yifan Jiang,Zhangyang Wang

Deep Neural Networks (DNNs) are well-known to act as over-parameterized deep ima ge priors (DIP) that regularize various image inverse problems. Meanwhile, researchers also proposed extremely compact, under-parameterized image priors (e.g., deep decoder) that are strikingly competent for image restoration too, despite a loss of accuracy. These two extremes push us to think whether there exists a better solution in the middle: between over- and under-parameterized image priors, can one identify "intermediate" parameterized image priors that achieve better trade-offs between performance, efficiency, and even preserving strong transferability? Drawing inspirations from the lottery ticket hypothesis (LTH), we conjecture and study a novel "lottery image prior" (LIP) by exploiting DNN inherent sparsity, stated as: given an over-parameterized DNN-based image prior, it will contain a sparse subnetwork that can be trained in isolation, to match the original DNN's performance when being applied as a prior to various image inverse problems}. Our results validate the superiority of LIPs: we can successfully locate the LIP subnetworks from over-parameterized DIPs at substantial sparsity ranges. Those LIP subnetworks significantly outperform deep decoders under comparably compact model sizes (by often fully preserving the effectiveness of their over-parameterized counterparts), and they also possess high transferability across different images as well as restoration task types. Besides, we also extend LIP to compressive sensing image reconstruction, where a pre-trained GAN generator is used as the prior (in contrast to untrained DIP or deep decoder), and confirm its validity in this setting too. To our best knowledge, this is the first time that LTH is demonstrated to be relevant in the context of inverse problems or image priors. Codes will be publicly available upon acceptance.

****

Generalizable Person Re-identification Without Demographics

YiFan Zhang,Feng Li,Zhang Zhang,Baosheng Yu,Liang Wang,Dacheng Tao,Tieniu Tan

Domain generalizable person re-identification (DG-ReID) aims to learn a ready-to-use domain-agnostic model directly for cross-dataset/domain evaluation, while current methods mainly explore the demographic information such as domain and/or camera labels for domain-invariant representation learning. However, the above-mentioned demographic information is not always accessible in practice due to privacy and security issues. In this paper, we consider the problem of person re-identification in a more general setting, \ie domain generalizable person re-identification without demographics (\textbf{DGWD-ReID}). To address the underlying uncertainty of domain distribution, we introduce distributionally robust optimization (DRO) to learn robust person re-identification models that perform well on all possible data distributions within the uncertainty set without demographics. However, directly applying the popular Kullback-Leibler divergence constrained DRO (or KL-DRO) fails to generalize well under the distribution shifts in real-world scenarios, since the convex condition may not hold for overparameterized neural networks. Inspired by this, we analyze and reformulate the popular KL-DRO by applying the change-of-measure technique, and then propose a simple yet efficient approach, \textbf{Unit-DRO}, which minimizes the loss over a new dataset with hard samples upweighted and other samples downweighted. We perform extensive experiments on both domain generalizable and cross-domain person re-identification tasks, and the empirical results on several large-scale benchmarks show that \iw~achieves superior performance compared to all baselines without using demographics.

****

LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation

Xuheng Cai,Chao Huang,Lianghao Xia,Xubin Ren

Graph neural network (GNN) is a powerful learning approach for graph-based recommender systems. Recently, GNNs integrated with contrastive learning have shown superior performance in recommendation with their data augmentation schemes, aiming at dealing with highly sparse data. Despite their success, most existing graph contrastive learning methods either perform stochastic augmentation (e.g., node/edge perturbation) on the user-item interaction graph, or rely on the heuristic-based augmentation techniques (e.g., user clustering) for generating contrastive views. We argue that these methods cannot well preserve the intrinsic semanti c structures and are easily biased by the noise perturbation. In this paper, we propose a simple yet effective graph contrastive learning paradigm LightGCL that mitigates these issues impairing the generality and robustness of CL-based recommenders. Our model exclusively utilizes singular value decomposition for contrastive augmentation, which enables the unconstrained structural refinement with global collaborative relation modeling. Experiments conducted on several benchmark datasets demonstrate the significant improvement in performance of our model over the state-of-the-arts. Further analyses demonstrate the superiority of LightGCL's robustness against data sparsity and popularity bias. The source code of our model is available at https://github.com/HKUDS/LightGCL.

****

MemoNav: Working Memory Model for Visual Navigation

Hongxin Li,Xu Yang,Zeyu Wang,yuran Yang,Shuqi Mei,Zhaoxiang Zhang

We present MemoNav, a novel memory model for image-goal navigation, which utilizes a working memory-inspired pipeline to improve navigation performance. Specifically, the node features on the topological map are stored in the short-term memory (STM), as these features are dynamically updated. The MemoNav retains the informative fraction of the STM via a forgetting module to improve navigation efficiency. To learn a global representation of 3D scenes, we introduce long-term memory (LTM) that continuously aggregates the STM. Afterward, a graph attention module encodes the retained STM and the LTM to generate working memory (WM). After encoding, the WM contains the informative features in the retained STM and the scene-level feature in the LTM and is finally used to generate actions. Consequently, the synergy of these three types of memory increases navigation performance by selectively retaining goal-relevant information and learning a high-level scene feature. When evaluated on multi-goal tasks, the MemoNav outperforms the SoTA methods at all difficulty levels in both Gibson and Matterport3D scenes. The MemoNav also achieves consistent improvements on traditional 1-goal tasks. Moreover, the qualitative results show that our model is less likely to be trapped in a deadlock.

****

Write and Paint: Generative Vision-Language Models are Unified Modal Learners

Shizhe Diao,Wangchunshu Zhou,Xinsong Zhang,Jiawei Wang

Recent advances in vision-language pre-training have pushed the state-of-the-art on various vision-language tasks, making machines more capable of multi-modal writing (image-to-text generation) and painting (text-to-image generation). However, few studies investigate if these two essential capabilities can be learned together and boost each other, making a versatile and powerful multi-modal foundation model. In this work, we disclose the potential of symmetric generative vision-language pre-training in learning to write and paint concurrently, and propose a new unified modal model, named DaVinci, trained with prefix language modeling and prefix image modeling, a simple generative self-supervised objective on image-text pairs. Thanks to the proposed prefix multi-modal modeling framework, DaVinci is simple to train, scalable to huge data, adaptable to both writing and painting tasks, and also strong on other vision, text, and multi-modal understanding tasks. DaVinci achieves competitive performance on a wide range of 27 generation/understanding tasks and demonstrates the superiority of combining vision/language generative pre-training. Furthermore, we carefully benchmark the performance of different vision-language pre-training objectives on different scales of pre-training datasets on a heterogeneous and broad distribution coverage. Our results demonstrate the potential of exploiting self-supervision in both language and vision inputs, and establish new, stronger baselines for future comparisons at different data scales. The code and pre-trained models are available at https://github.com/shizhediao/DaVinci.

****

Progressive Voronoi Diagram Subdivision Enables Accurate Data-free Class-Incremental Learning

Chunwei Ma,Zhanghexuan Ji,Ziyun Huang,Yan Shen,Mingchen Gao,Jinhui Xu

Data-free Class-incremental Learning (CIL) is a challenging problem because rehearsing data from previous phases is strictly prohibited, causing catastrophic fo rgetting of Deep Neural Networks (DNNs). In this paper, we present \emph{iVoro}, a novel framework derived from computational geometry. We found Voronoi Diagram (VD), a classical model for space subdivision, is especially powerful for solving the CIL problem, because VD itself can be constructed favorably in an incremental manner -- the newly added sites (classes) will only affect the proximate classes, making the non-contiguous classes hardly forgettable. Furthermore, we bridge DNN and VD using Power Diagram Reduction, and show that the VD structure can be progressively refined along the phases using a divide-and-conquer algorithm. Moreover, our VD construction is not restricted to the deep feature space, but is also applicable to multiple intermediate feature spaces, promoting VD to be multilayer VD that efficiently captures multi-grained features from DNN. Importantly, \emph{iVoro} is also capable of handling uncertainty-aware test-time Voronoi cell assignment and has exhibited high correlations between geometric uncertainty and predictive accuracy (up to ${\sim}0.9$). Putting everything together, \emph{iVoro} achieves up to 25.26%, 37.09%, and 33.21% improvements on CIFAR-100, TinyImageNet, and ImageNet-Subset, respectively, compared to the state-of-the-art non-exemplar CIL approaches. In conclusion, \emph{iVoro} enables highly accurate, privacy-preserving, and geometrically interpretable CIL that is particularly useful when cross-phase data sharing is forbidden, e.g. in medical applications.

****

Data Valuation Without Training of a Model

Ki Nohyun,Hoyong Choi,Hye Won Chung

Many recent works on understanding deep learning try to quantify how much individual data instances influence the optimization and generalization of a model. Such attempts reveal characteristics and importance of individual instances, which may provide useful information in diagnosing and improving deep learning. However, most of the existing works on data valuation require actual training of a model, which often demands high-computational cost. In this paper, we provide a training-free data valuation score, called complexity-gap score, which is a data-centric score to quantify the influence of individual instances in generalization of two-layer overparameterized neural networks. The proposed score can quantify irregularity of the instances and measure how much each data instance contributes in the total movement of the network parameters during training. We theoretically analyze and empirically demonstrate the effectiveness of the complexity-gap score in finding `irregular or mislabeled' data instances, and also provide applications of the score in analyzing datasets and diagnosing training dynamics. Our code is publicly available at https://github.com/JJchy/CG_score.

****

HotProtein: A Novel Framework for Protein Thermostability Prediction and Editing

Tianlong Chen,Chengyue Gong,Daniel Jesus Diaz,Xuxi Chen,Jordan Tyler Wells,qiang liu,Zhangyang Wang,Andrew Ellington,Alex Dimakis,Adam Klivans

The molecular basis of protein thermal stability is only partially understood and has major significance for drug and vaccine discovery. The lack of datasets and standardized benchmarks considerably limits learning-based discovery methods. We present \texttt{HotProtein}, a large-scale protein dataset with \textit{growth temperature} annotations of thermostability, containing 182K amino acid sequences and 3K folded structures from 230 different species with a wide temperature range $-20^{\circ}\texttt{C}\sim 120^{\circ}\texttt{C}$. Due to functional domain differences and data scarcity within each species, existing methods fail to generalize well on our dataset. We address this problem through a novel learning framework, consisting of (1) Protein structure-aware pre-training (SAP) which leverages 3D information to enhance sequence-based pre-training; (2) Factorized sparse tuning (FST) that utilizes low-rank and sparse priors as an implicit regularization, together with feature augmentations. Extensive empirical studies demonstrate that our framework improves thermostability prediction compared to other deep learning models. Finally, we introduce a novel editing algorithm to efficiently generate positive amino acid mutations that improve thermostability. Codes are available in https://github.com/VITA-Group/HotProtein.

****

Agent-Controller Representations: Principled Offline RL with Rich Exogenous Information

Riashat Islam,Manan Tomar,Alex Lamb,Hongyu Zang,Yonathan Efroni,Dipendra Misra,Xin Li,Harm van Seijen,Remi Tachet des Combes,John Langford

Learning to control an agent from data collected offline in a rich pixel-based visual observation space is vital for real-world applications of reinforcement learning (RL). A major challenge in this setting is the presence of input information that is hard to model and irrelevant to controlling the agent. This problem has been approached by the theoretical RL community through the lens of exogenous information, i.e, any control-irrelevant information contained in observations. For example, a robot navigating in busy streets needs to ignore irrelevant information, such as other people walking in the background, textures of objects, or birds in the sky. In this paper, we focus on the setting with visually detailed exogenous information, and introduce new offline RL benchmarks offering the ability to study this problem. We find that contemporary representation learning techniques can fail on datasets where the noise is a complex and time dependent process, which is prevalent in practical applications. To address these, we propose to use multi-step inverse models, which have seen a great deal of interest in the RL theory community, to learn Agent-Controller Representations for Offline-RL (ACRO). Despite being simple and requiring no reward, we show theoretically and empirically that the representation created by this objective greatly outperforms baselines.

****

RPM: Generalizable Multi-Agent Policies for Multi-Agent Reinforcement Learning

Wei Qiu,Xiao Ma,Bo An,Svetlana Obraztsova,Shuicheng YAN,Zhongwen Xu

Despite the recent advancement in multi-agent reinforcement learning (MARL), the MARL agents easily overfit the training environment and perform poorly in evaluation scenarios where other agents behave differently. Obtaining generalizable policies for MARL agents is thus necessary but challenging mainly due to complex multi-agent interactions. In this work, we model the MARL problem with Markov Games and propose a simple yet effective method, called ranked policy memory (RPM), i.e., to maintain a look-up memory of policies to achieve good generalizability. The main idea of RPM is to train MARL policies via gathering massive multi-agent interaction data. In particular, we first rank each agent’s policies by its training episode return, i.e., the episode return of each agent in the training environment; we then save the ranked policies in the memory; when an episode starts, each agent can randomly select a policy from the RPM as the behavior policy. Each agent uses the behavior policy to gather multi-agent interaction data for MARL training. This innovative self-play framework guarantees the diversity of multi-agent interaction in the training data. Experimental results on Melting Pot demonstrate that RPM enables MARL agents to interact with unseen agents in multi-agent generalization evaluation scenarios and complete given tasks. It significantly boosts the performance up to 818% on average.

****

Behavior Prior Representation learning for Offline Reinforcement Learning

Hongyu Zang,Xin Li,Jie Yu,Chen Liu,Riashat Islam,Remi Tachet des Combes,Romain Laroche

Offline reinforcement learning (RL) struggles in environments with rich and noisy inputs, where the agent only has access to a fixed dataset without environment interactions. Past works have proposed common workarounds based on the pre-training of state representations, followed by policy training. In this work, we introduce a simple, yet effective approach for learning state representations. Our method, Behavior Prior Representation (BPR), learns state representations with an easy-to-integrate objective based on behavior cloning of the dataset: we first learn a state representation by mimicking actions from the dataset, and then train a policy on top of the fixed representation, using any off-the-shelf Offline RL algorithm. Theoretically, we prove that BPR carries out performance guarantees when integrated into algorithms that have either policy improvement guarantees (conservative algorithms) or produce lower bounds of the policy values (pessimistic algorithms). Empirically, we show that BPR combined with existing state-of

-the-art Offline RL algorithms leads to significant improvements across several offline control benchmarks. The code is available at \url{https://github.com/bit1029public/offline_bpr}

****

How Does Adaptive Optimization Impact Local Neural Network Geometry?

Kaiqi Jiang,Dhruv Malik,Yuanzhi Li

Adaptive optimization methods are well known to achieve superior convergence relative to vanilla gradient methods. The traditional viewpoint in optimization, particularly in convex optimization, explains this improved performance by arguing that, unlike vanilla gradient schemes, adaptive algorithms mimic the behavior of a second-order method by adapting to the global geometry of the loss function. We argue that in the context of neural network optimization, this traditional viewpoint is insufficient. Instead, we advocate for a local trajectory analysis. For iterate trajectories produced by running a generic optimization algorithm OPT, we introduce $R^{\text{OPT}}_{\text{med}}$, a statistic that is analogous to the condition number of the loss Hessian evaluated at the iterates. Through extensive experiments, we show that adaptive methods such as Adam bias the trajectories towards regions where $R^{\text{Adam}}_{\text{med}}$ is small, where one might expect faster convergence. By contrast, vanilla gradient methods like SGD bias the trajectories towards regions where $R^{\text{SGD}}_{\text{med}}$ is comparatively large. We complement these empirical observations with a theoretical result that provably demonstrates this phenomenon in the simplified setting of a two-layer linear network. We view our findings as evidence for the need of a new explanation of the success of adaptive methods, one that is different than the conventional wisdom.

****

Substructured Graph Convolution for Non-overlapping Graph Decomposition

Youngkyu Lee,Chang-Ock Lee

Graph convolutional networks have been widely used to solve the graph problems such as node classification, link prediction, and recommender systems. It is well known that large graphs require large amount of memory and time to train graph convolutional networks. To deal with large graphs, many methods are being done, such as graph sampling or decomposition. In particular, graph decomposition has the advantage of parallel computation, but information loss occurs in the interface part. In this paper, we propose a novel substructured graph convolution that reinforces the interface part lost by graph decomposition. Numerical results indicate that the proposed method is robust in the number of subgraphs compared to other methods.

****

On the Neural Tangent Kernel of Equilibrium Models

Zhili Feng,J Zico Kolter

This work studies the neural tangent kernel (NTK) of deep equilibrium (DEQ) model, a practical ``infinite-depth'' architecture which directly computes the infinite-depth limit of a weight-tied network via root-finding. Even though the NTK of a fully-connected neural network is stochastic if its width and depth both tend to infinity simultaneously, we show that contrarily a DEQ model still enjoys a deterministic NTK despite its width and depth going to infinity at the same time. Moreover, such deterministic NTK can be found efficiently via root-finding.

****

From Play to Policy: Conditional Behavior Generation from Uncurated Robot Data

Zichen Jeff Cui,Yibin Wang,Nur Muhammad Mahi Shafiullah,Lerrel Pinto

While large-scale sequence modelling from offline data has led to impressive performance gains in natural language generation and image generation, directly translating such ideas to robotics has been challenging. One critical reason for this is that uncurated robot demonstration data, i.e. play data, collected from non-expert human demonstrators are often noisy, diverse, and distributionally multi-modal. This makes extracting useful, task-centric behaviors from such data a difficult generative modelling problem. In this work, we present Conditional Behavior Transformers (C-BeT), a method that combines the multi-modal generation ability of Behavior Transformer with future-conditioned goal specification. On a su ite of simulated benchmark tasks, we find that C-BeT improves upon prior state-of-the-art work in learning from play data by an average of 45.7%. Further, we demonstrate for the first time that useful task-centric behaviors can be learned on a real-world robot purely from play data without any task labels or reward information. Robot videos are best viewed on our project website: play-to-policy.github.io

****

Beyond Counting Linear Regions of Neural Networks, Simple Linear Regions Dominate!

FENGLEI FAN,Wei Huang,Lecheng Ruan,Tieyong Zeng,Huan Xiong,Fei Wang

Functions represented by a neural network with the widely-used ReLU activation are piecewise linear functions over linear regions (polytopes). Figuring out the properties of such polytopes is of fundamental importance for the development of neural networks.

So far, either theoretical or empirical studies on polytopes stay at the level of counting their number. Despite successes in explaining the power of depth and so on, counting the number of polytopes puts all polytopes on an equal booting, which is essentially an incomplete characterization of polytopes. Beyond counting, here we study the shapes of polytopes via the number of simplices obtained by triangulations of polytopes. First, we demonstrate the properties of the number of simplices in triangulations of polytopes, and compute the upper and lower bounds of the maximal number of simplices that a network can generate. Next, by computing and analyzing the histogram of simplices across polytopes, we find that a ReLU network has surprisingly uniform and simple polytopes, although these polytopes theoretically can be rather diverse and complicated. This finding is a novel implicit bias that concretely reveals what kind of simple functions a network learns and sheds light on why deep learning does not overfit. Lastly, we establish a theorem to illustrate why polytopes produced by a deep network are simple and uniform. The core idea of the proof is counter-intuitive: adding depth probably does not create a more complicated polytope. We hope our work can inspire more research into investigating polytopes of a ReLU neural network, thereby upgrading the knowledge of neural networks to a new level.

****

Recovering Top-Two Answers and Confusion Probability in Multi-Choice Crowdsourcing

Hyeonsu Jeong,Hye Won Chung

We consider multi-choice crowdsourced labeling with the goal of recovering not only the ground truth but also the most confusing answer and the confusion probability. The most confusing answer provides useful information about the task by revealing the most plausible answer other than the ground truth and how plausible it is. To theoretically analyze such scenarios, we propose a model where there are top-two plausible answers for each task, distinguished from the rest of choices. Task difficulty is quantified by the confusion probability between the top two, and worker reliability is quantified by the probability of giving an answer among the top two. Under this model, we propose a two-stage inference algorithm to infer the top-two answers, where the first stage uses the spectral method to obtain an initial estimate for the top two, and the second stage uses the result of the first stage to refine the estimates based on the maximum likelihood estimator (MLE). We show that our algorithm achieves the minimax optimal convergence rate. We conduct both synthetic and real-data experiments and demonstrate that our algorithm achieves the performance near the optimal MLE for synthetic datasets and the best performance for real datasets compared to other recent algorithms. This shows that our model explains well the real datasets with heterogeneous task difficulties due to confusion between plausible answers.

****

SCALE-UP: An Efficient Black-box Input-level Backdoor Detection via Analyzing Scaled Prediction Consistency

Junfeng Guo,Yiming Li,Xun Chen,Hanqing Guo,Lichao Sun,Cong Liu

Deep neural networks (DNNs) are vulnerable to backdoor attacks, where adversarie

s embed a hidden backdoor trigger during the training process for malicious prediction manipulation. These attacks pose great threats to the applications of DNNs under the real-world machine learning as a service (MLaaS) setting, where the deployed model is fully black-box while the users can only query and obtain its predictions. Currently, there are many existing defenses to reduce backdoor threats. However, almost all of them cannot be adopted in MLaaS scenarios since they require getting access to or even modifying the suspicious models. In this paper, we propose a simple yet effective black-box input-level backdoor detection, called SCALE-UP, which requires only the predicted labels to alleviate this problem. Specifically, we identify and filter malicious testing samples by analyzing their prediction consistency during the pixel-wise amplification process. Our defense is motivated by an intriguing observation (dubbed \emph{scaled prediction consistency}) that the predictions of poisoned samples are significantly more consistent compared to those of benign ones when amplifying all pixel values. Besides, we also provide theoretical foundations to explain this phenomenon. Extensive experiments are conducted on benchmark datasets, verifying the effectiveness and efficiency of our defense and its resistance to potential adaptive attacks. Our codes are available at \url{https://github.com/JunfengGo/SCALE-UP}.
****

On the Perils of Cascading Robust Classifiers
Ravi Mangal,Zifan Wang,Chi Zhang,Klas Leino,Corina Pasareanu,Matt Fredrikson
Ensembling certifiably robust neural networks is a promising approach for improving the \emph{certified robust accuracy} of neural models.
Black-box ensembles that assume only query-access to the constituent models (and their robustness certifiers) during prediction are particularly attractive due to their modular structure. Cascading ensembles are a popular instance of black-box ensembles that appear to improve certified robust accuracies in practice. However, we show that the robustness certifier used by a cascading ensemble is unsound. That is, when a cascading ensemble is certified as locally robust at an input x (with respect to ϵ), there can be inputs x' in the ϵ-ball centered at x, such that the cascade's prediction at x' is different from x and thus the ensemble is not locally robust. Our theoretical findings are accompanied by empirical results that further demonstrate this unsoundness. We present a new attack against cascading ensembles and show that: (1) there exists an adversarial input for up to 88\% of the samples where the ensemble claims to be certifiably robust and accurate; and (2) the accuracy of a cascading ensemble under our attack is as low as 11\% when it claims to be certifiably robust and accurate on 97\% of the test set. Our work reveals a critical pitfall of cascading certifiably robust models by showing that the seemingly beneficial strategy of cascading can actually hurt the robustness of the resulting ensemble. Our code is available at https://github.com/TristaChi/ensembleKW.
****

Visual Classification via Description from Large Language Models
Sachit Menon,Carl Vondrick
Vision-language models such as CLIP have shown promising performance on a variety of recognition tasks using the standard zero-shot classification procedure -- computing similarity between the query image and the embedded words for each category. By only using the category name, they neglect to make use of the rich context of additional information that language affords. The procedure gives no intermediate understanding of why a category is chosen, and furthermore provides no mechanism for adjusting the criteria used towards this decision. We present an alternative framework for classification with VLMs, which we call classification by description. We ask VLMs to check for descriptive features rather than broad categories: to find a tiger, look for its stripes; its claws; and more. By basing decisions on these descriptors, we can provide additional cues that encourage using the features we want to be used. In the process, we can get a clear idea of what the model ``thinks" it is seeing to make its decision; it gains some level of inherent explainability. We query large language models (e.g., GPT-3) for these descriptors to obtain them in a scalable way. Extensive experiments show our framework has numerous advantages past interpretability. We show improvements

in accuracy on ImageNet across distribution shifts; demonstrate the ability to adapt VLMs to recognize concepts unseen during training; and illustrate how descriptors can be edited to effectively mitigate bias compared to the baseline.

****

Contrastive Novelty Learning: Anticipating Outliers with Large Language Models

Albert Xu,Xiang Ren,Robin Jia

In many task settings, text classification models are likely to encounter examples from novel classes on which they cannot predict correctly. Selective prediction, in which models abstain on low-confidence examples, provides a possible solution, but existing models are often overly confident on OOD examples. To remedy this overconfidence, we introduce Contrastive Novelty Learning (CNL), a two-step method that generates OOD examples representative of novel classes, then trains to decrease confidence on them. First, we generate OOD examples by prompting a large language model twice: we prompt it to enumerate novel classes relevant to the label set, then generate examples from each novel class matching the task format. Second, we train our classifier with a novel contrastive objective that encourages lower confidence on generated OOD examples than training examples. When trained with CNL, classifiers improve in their ability to detect and abstain on OOD examples over prior methods by an average of 2.3% AUAC and 5.5% AUROC across 4 NLP datasets, with no cost to in-distribution accuracy.

****

MIMT: Masked Image Modeling Transformer for Video Compression

Jinxi Xiang,Kuan Tian,Jun Zhang

Deep learning video compression outperforms its hand-craft counterparts with enhanced flexibility and capacity. One key component of the learned video codec is the autoregressive entropy model conditioned on spatial and temporal priors. Operating autoregressive on raster scanning order naively treats the context as unidirectional. This is neither efficient nor optimal, considering that conditional information probably locates at the end of the sequence. We thus introduce an entropy model based on a masked image modeling transformer (MIMT) to learn the spatial-temporal dependencies. Video frames are first encoded into sequences of tokens and then processed with the transformer encoder as priors. The transformer decoder learns the probability mass functions (PMFs) \emph{conditioned} on the priors and masked inputs. Then it is capable of selecting optimal decoding orders without a fixed direction. During training, MIMT aims to predict the PMFs of randomly masked tokens by attending to tokens in all directions. This allows MIMT to capture the temporal dependencies from encoded priors and the spatial dependencies from the unmasked tokens, i.e., decoded tokens. At inference time, the model begins with generating PMFs of all masked tokens in parallel and then decodes the frame iteratively from the previously-selected decoded tokens (i.e., with high confidence). In addition, we improve the overall performance with more techniques, e.g., manifold conditional priors accumulating a long range of information, shifted window attention to reduce complexity. Extensive experiments demonstrate the proposed MIMT framework equipped with the new transformer entropy model achieves state-of-the-art performance on HEVC, UVG, and MCL-JCV datasets, generally outperforming the VVC in terms of PSNR and SSIM.

****

Speculative Decoding: Lossless Speedup of Autoregressive Translation

Heming Xia,Tao Ge,Si-Qing Chen,Furu Wei,Zhifang Sui

Different from some previous work accelerating autoregressive translation (AT) at the sacrifice of quality, we propose Speculative Decoding (SpecDec) -- a novel decoding paradigm inspired by speculative execution in computer architecture, which combines respective advantages of AT and non-autoregressive translation (NAT) for lossless speedup of translation. At each decoding step, SpecDec first speculatively drafts (i.e. decodes) next k tokens with an NAT model and then verifies them with an AT model, where only the drafted tokens passing the verification are accepted as decoded tokens for guaranteeing its translation result is exactly the same as AT. The collaboration of NAT drafting and AT verification leads to a much higher decoding speed without quality loss due to parallel computing enabled by speculative decoding.

We conduct experiments in 4 standard WMT translation benchmarks and confirm the vanilla SpecDec yields exactly the same results as AT greedy decoding with an around $3\times$ speedup, and that its variant (SpecDec++) with an advanced verification strategy not only outperforms AT greedy decoding, but also further improves the decoding speed, resulting in an around $5\times$ speedup over AT. Moreover, SpecDec can be easily generalized for speeding up other seq2seq tasks like Abstractive Summarization, and benefit more from stronger computing devices, demonstrating its potential to become a de facto decoding standard in the future for efficient and lossless seq2seq generation.
****

$$CONVOLUTION AND POOLING OPERATION MODULE WITH ADAPTIVE STRIDE PROCESSING EFFEC$$

■■ ■,■■ ■

$$Convolutional neural network is one of the representative models of deep learning, which has a wide range of applications. Convolution and pooling are two key operations in convolutional neural networks. They play an important role in extracting input features and mapping low-level semantic features to high-level semantic features. Stride is an important parameter involved in convolution and pooling operations, which refers to the distance of each slide of the convolution kernel (pooling kernel) during the convolution (pooling) operation. The stride has an impact on the granularity of feature extraction and the selection (filtering) of features, thus affecting the performance of convolutional neural networks. At present, in the training of convolutional neural networks, the content of convolution kernel and pooling kernel can be determined by the optimization algorithm based on gradient descent. However, the stride usually cannot be treated similarly, and can only be selected manually as a hyperparameter. Most of the existing related works choose a fixed stride, for example, the value is 1. In fact, different tasks or inputs may require different stride for better model processing. Therefore, this paper views the role of stride in convolution and pooling operation from the perspective of sampling, and proposes a convolution and pooling operation module with adaptive stride processing effect. The feature of the proposed module is that the feature map finally obtained by convolution or pooling operation is no longer limited to equal interval downsampling (feature extraction) according to a fixed stride, but adaptively extracted according to the changes of input features. We apply the proposed module on many convolutional neural network models, including VGG, Alexnet and MobileNet for image classification, YOLOX-S for object detection, Unet for image segmentation, and so on. Simulation results show that the proposed module can effectively improve the perform$$
****

Transformer Module Networks for Systematic Generalization in Visual Question Answering

Moyuru Yamada,Vanessa D'Amario,Kentaro Takemoto,Xavier Boix,Tomotake Sasaki

Transformers achieve great performance on Visual Question Answering (VQA). However, their systematic generalization capabilities, i.e., handling novel combinations of known concepts, is unclear. We reveal that Neural Module Networks (NMNs),

i.e., question-specific compositions of modules that tackle a sub-task, achieve better or similar systematic generalization performance than the conventional Transformers, even though NMNs' modules are CNN-based. In order to address this shortcoming of Transformers with respect to NMNs, in this paper we investigate whether and how modularity can bring benefits to Transformers. Namely, we introduce Transformer Module Network (TMN), a novel NMN based on compositions of Transformer modules. TMNs achieve state-of-the-art systematic generalization performance in three VQA datasets, improving more than 30% over standard Transformers for novel compositions of sub-tasks. We show that not only the module composition but also the module specialization for each sub-task are the key of such performance gain.
****

Diving into Unified Data-Model Sparsity for Class-Imbalanced Graph Representation Learning

Chunhui Zhang,Chao Huang,Yijun Tian,Qianlong Wen,Zhongyu Ouyang,Youhuan Li,Yanfang Ye,Chuxu Zhang

Even pruned by the state-of-the-art network compression methods, recent research shows that deep learning model training still suffers from the demand of massive data usage. In particular, Graph Neural Networks (GNNs) trained upon non-Euclidean graph data often encounter relatively higher time costs, due to its irregular and nasty density properties, compared with data in the regular Euclidean space (e.g., image or text). Another natural property accompanied with graphs is class-imbalance which cannot be alleviated even with massive data, therefore hinders GNNs’ ability in generalization. To fully tackle these unpleasant properties, (i) theoretically, we introduce a hypothesis about to what extent a subset of the training data can approximate the full dataset’s learning effectiveness. The effectiveness is further guaranteed and proved by the gradients’ distance between the subset and the full set; (ii) empirically, we discover that during the learning process of a GNN, some samples in the training dataset are informative in providing gradients to update model parameters. Moreover, the informative subset evolves dynamically during the training process, for samples that are informative in the current training epoch may not be so in the next one. We refer this observation as dynamic data sparsity. We also notice that sparse subnets pruned from a well-trained GNN sometimes forget the information provided by the informative subset, reflected in their poor performance upon the subset. Based on these findings, we develop a unified data-model dynamic sparsity framework named Graph Decantation (GraphDec) to address challenges brought by training upon a massive class-imbalanced graph dataset. The key idea of GraphDec is to identify the informative subset dynamically during the training process by adopting sparse graph contrastive learning. Extensive experiments on multiple benchmark datasets demonstrate that GraphDec outperforms state-of-the-art baselines for the class-imbalanced graph classification and class-imbalanced node classification tasks, with respect to classification accuracy and data usage efficiency.
****

Learning Math Reasoning from Self-Sampled Correct and Partially-Correct Solutions

Ansong Ni,Jeevana Priya Inala,Chenglong Wang,Alex Polozov,Christopher Meek,Dragomir Radev,Jianfeng Gao

Pretrained language models have shown superior performance on many natural language processing tasks, yet they still struggle at multi-step formal reasoning tasks like grade school math problems. One key challenge of finetuning them to solve such math reasoning problems is that many existing datasets only contain one reference solution for each problem, despite the fact that there are often alternative solutions resembling different reasoning paths to the final answer. This way, the finetuned models are biased towards the limited reference solutions, which limits their generalization to unseen examples. To mitigate this issue, we propose to let the model perform sampling during training and learn from both self-sampled fully-correct solutions, which yield the correct answer upon execution, and partially-correct solutions, whose intermediate state matches an intermediate state of a known correct solution. We show that our use of self-sampled corre ct and partially-correct solutions can benefit learning and help guide the sampling process, leading to more efficient exploration of the solution space. Additionally, we explore various training objectives to support learning from multiple solutions per example and find they greatly affect the performance. Experiments on two math reasoning datasets show the effectiveness of our method compared to learning from a single reference solution with MLE, where we improve PASS@100 from 35.5% to 44.5% for GSM8K, and 27.6% to 36.2% PASS@80 for MathQA. Such improvements are also consistent across different model sizes.

****

Adaptive Robust Evidential Optimization For Open Set Detection from Imbalanced Data

Hitesh Sapkota,Qi Yu

Open set detection (OSD) aims at identifying data samples of an unknown class ($i.e.$, open set) from those of known classes ($i.e.$, closed set) based on a model trained from closed set samples. However, a closed set may involve a highly imbalanced class distribution. Accurately differentiating open set samples and those from a minority class in the closed set poses a fundamental challenge as the model may be equally uncertain when recognizing samples from the minority class. In this paper, we propose Adaptive Robust Evidential Optimization (AREO) that offers a principled way to quantify sample uncertainty through evidential learning while optimally balancing the model training over all classes in the closed set through adaptive distributively robust optimization (DRO). To avoid the model to primarily focus on the most difficult samples by following the standard DRO, adaptive DRO training is performed, which is governed by a novel multi-scheduler learning mechanism to ensure an optimal model training behavior that gives sufficient attention to the difficult samples and the minority class while capable of learning common patterns from the majority classes. Our experimental results on multiple real-world datasets demonstrate that the proposed model outputs uncertainty scores that can clearly separate samples from closed and open sets, respectively, and the detection results outperform the competitive baselines.

****

The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation

Zihui Xue,Zhengqi Gao,Sucheng Ren,Hang Zhao

Crossmodal knowledge distillation (KD) extends traditional knowledge distillation to the area of multimodal learning and demonstrates great success in various applications. To achieve knowledge transfer across modalities, a pretrained network from one modality is adopted as the teacher to provide supervision signals to a student network learning from the other modality. In contrast to the empirical success reported in prior works, the working mechanism of crossmodal KD remains a mystery. In this paper, we present a thorough understanding of crossmodal KD. We begin by providing two failure cases and demonstrate that KD is not a universal cure in crossmodal knowledge transfer. We then present the modality Venn diagram to understand modality relationships and the modality focusing hypothesis revealing the decisive factor in the efficacy of crossmodal KD. Experimental results on 6 multimodal datasets help justify our hypothesis, diagnose failure cases, and point directions to improve crossmodal knowledge transfer in the future.

****

Hungry Hungry Hippos: Towards Language Modeling with State Space Models

Daniel Y Fu,Tri Dao,Khaled Kamal Saab,Armin W Thomas,Atri Rudra,Christopher Re

State space models (SSMs) have demonstrated state-of-the-art sequence modeling performance in some modalities, but underperform attention in language modeling. Moreover, despite scaling nearly linearly in sequence length instead of quadratically, SSMs are still slower than Transformers due to poor hardware utilization. In this paper, we make progress on understanding the expressivity gap between SSMs and attention in language modeling, and on reducing the hardware barrier between SSMs and attention. First, we use synthetic language modeling tasks to understand the gap between SSMs and attention. We find that existing SSMs struggle with two capabilities: recalling earlier tokens in the sequence and comparing tokens across the sequence. To understand the impact on language modeling, we propo se a new SSM layer, H3, that is explicitly designed for these abilities. H3 matches attention on the synthetic languages and comes within 0.4 PPL of Transformers on OpenWebText. Furthermore, a hybrid 125M-parameter H3-attention model that retains two attention layers surprisingly outperforms Transformers on OpenWebText by 1.0 PPL. Next, to improve the efficiency of training SSMs on modern hardware, we propose FlashConv. FlashConv uses a fused block FFT algorithm to improve efficiency on sequences up to 8K, and introduces a novel state passing algorithm that exploits the recurrent properties of SSMs to scale to longer sequences. FlashConv yields 2$\times$ speedup on the long-range arena benchmark and allows hybrid language models to generate text 2.4$\times$ faster than Transformers. Using FlashConv, we scale hybrid H3-attention language models up to 2.7B parameters on the Pile and find promising initial results, achieving lower perplexity than Transformers and outperforming Transformers in zero- and few-shot learning on a majority of tasks in the SuperGLUE benchmark.

****

FINE: Future-Aware Inference for Streaming Speech Translation

Biao Fu,Kai Fan,Minpeng Liao,Zhongqiang Huang,Boxing Chen,Xiaodong Shi,Yidong Chen

A popular approach to streaming speech translation is to employ a single offline model together with a \textit{wait-k} policy to support different latency requirements. It is a simpler alternative compared to training multiple online models with different latency constraints. However, there is an apparent mismatch in using a model trained with complete utterances on partial streaming speech during online inference. We demonstrate that there is a significant difference between the speech representations extracted at the end of a streaming input and their counterparts at the same positions when the complete utterance is available. Built upon our observation that this problem can be alleviated by introducing a few frames of future speech signals, we propose \textbf{F}uture-aware \textbf{in}ferenc\textbf{e} (FINE) for streaming speech translation with two different methods to make the model aware of the future. The first method FINE-Mask incorporates future context through a trainable masked speech model. The second method FINE-Wait simply waits for more actual future audio frames at the cost of extra latency. Experiments on the MuST-C EnDe, EnEs and EnFr benchmarks show that both methods are effective and can achieve better trade-offs between translation quality and latency than strong baselines, and a hybrid approach combining the two can achieve further improvement. Extensive analyses suggest that our methods can effectively alleviate the aforementioned mismatch problem between offline training and online inference.

****

Dual Diffusion Implicit Bridges for Image-to-Image Translation

Xuan Su,Jiaming Song,Chenlin Meng,Stefano Ermon

Common image-to-image translation methods rely on joint training over data from both source and target domains. The training process requires concurrent access to both datasets, which hinders data separation and privacy protection; and existing models cannot be easily adapted for translation of new domain pairs. We present Dual Diffusion Implicit Bridges (DDIBs), an image translation method based on diffusion models, that circumvents training on domain pairs. Image translation with DDIBs relies on two diffusion models trained independently on each domain, and is a two-step process: DDIBs first obtain latent encodings for source images with the source diffusion model, and then decode such encodings using the target model to construct target images. Both steps are defined via ordinary differential equations (ODEs), thus the process is cycle consistent only up to discretization errors of the ODE solvers. Theoretically, we interpret DDIBs as concatenation of source to latent, and latent to target Schrodinger Bridges, a form of entropy-regularized optimal transport, to explain the efficacy of the method. Experimentally, we apply DDIBs on synthetic and high-resolution image datasets, to demonstrate their utility in a wide variety of translation tasks and their inherent optimal transport properties.

****

HYPERPRUNING: EFFICIENT PRUNING THROUGH LYAPUNOV METRIC HYPERSEARCH

Yang Zheng,Eli Shlizerman

A variety of pruning methods have been introduced for over-parameterized recurrent neural networks to improve efficiency in terms of power and storage. With the advance in pruning methods and their variety, a new problem of ‘hyperpruning’ is becoming apparent: finding a suitable pruning method with optimal hyperparameter configuration for a particular task and network. Such search is different from the standard hyperparameter search, where the accuracy of the optimal configuration is unknown. In the context of network pruning, the accuracy of the non-pruned (dense) model sets the target for the accuracy of the pruned model. Thereby, the goal of hyperpruning is to reach or even surpass this target. It is critical to develop efficient strategies for hyperpruning since direct search through pruned variants would require time-consuming training without guarantees for improved performance. To address this problem, we introduce a novel distance based on Lyapunov Spectrum (LS) which provides means to compare pruned variants with the dense model and early in training to estimate the accuracy that pruned variants will achieve after extensive training. The ability to predict performance allows us to incorporate the LS-based distance with Bayesian hyperparameter optimization methods and to propose an efficient and first-of-its-kind hyperpruning approach called LS-based Hyperpruning (LSH) which can optimize the search time by an order of magnitude compared to standard full training search with the loss (or perplexity) being the accuracy metric. Our experiments on stacked LSTM and RHN language models trained with the Penn Treebank dataset show that with a given budget of training epochs and desired pruning ratio, LSH obtains more optimal variants than standard loss-based hyperparameter optimization methods. Furthermore, as a result of the search, LSH identifies pruned variants that outperform state-of-the-art pruning methods and surpass the accuracy of the dense model.

****

Relational Curriculum Learning for Graph Neural Networks

Zheng Zhang,Junxiang Wang,Liang Zhao

Graph neural networks have achieved great success in representing structured data and its downstream tasks such as node classification. The key idea is to recursively propagate and aggregate information along the edges of a given graph topology. However, edges in real-world graphs often have varying degrees of difficulty, and some edges may even be noisy to the downstream tasks. Therefore, existing graph neural network models may lead to suboptimal learned representations because they usually consider every edge in a given graph topology equally. On the other hand, curriculum learning, which mimics the human learning principle of learning data samples in a meaningful order, has been shown to be effective in improving the generalization ability of representation learners by gradually proceeding from easy to more difficult samples during training. Unfortunately, most existing curriculum learning strategies are designed for i.i.d data samples and cannot be trivially generalized to handle structured data with dependencies. In order to address these issues, in this paper we propose a novel curriculum learning method for structured data to leverage the various underlying difficulties of data dependencies to improve the quality of learned representations on structured data. Specifically, we design a learning strategy that gradually incorporates edges in a given graph topology into training according to their difficulty from easy to hard, where the degree of difficulty is measured by a self-supervised learning paradigm. We demonstrate the strength of our proposed method in improving the generalization ability of learned representations through extensive experiments on nine synthetic datasets and seven real-world datasets with different commonly used graph neural network models as backbone models.

****

The World is Changing: Improving Fair Training under Correlation Shifts

Yuji Roh,Kangwook Lee,Steven Euijong Whang,Changho Suh

Model fairness is an essential element for Trustworthy AI. While many techniques for model fairness have been proposed, most of them assume that the training and deployment data distributions are identical, which is often not true in practice. In particular, when the bias between labels and sensitive groups changes, the fairness of the trained model is directly influenced and can worsen. We make t wo contributions for solving this problem. First, we analytically show that existing in-processing fair algorithms have fundamental limits in accuracy and fairness. We introduce the notion of correlation shifts, which can explicitly capture the change of the above bias. Second, we propose a novel pre-processing step that samples the input data to reduce correlation shifts and thus enables the in-processing approaches to overcome their limitations. We formulate an optimization problem for adjusting the data ratio among labels and sensitive groups to reflect the shifted correlation. A key advantage of our approach lies in decoupling the roles of pre-processing and in-processing approaches: correlation adjustment via pre-processing and unfairness mitigation on the processed data via in-processing. Experiments show that our framework effectively improves existing in-processing fair algorithms w.r.t. accuracy and fairness, both on synthetic and real datasets.

****

ACMP: Allen-Cahn Message Passing with Attractive and Repulsive Forces for Graph Neural Networks

Yuelin Wang,Kai Yi,Xinliang Liu,Yu Guang Wang,Shi Jin

Neural message passing is a basic feature extraction unit for graph-structured data considering neighboring node features in network propagation from one layer to the next. We model such process by an interacting particle system with attractive and repulsive forces and the Allen-Cahn force arising in the modeling of phase transition. The dynamics of the system is a reaction-diffusion process which can separate particles without blowing up. This induces an Allen-Cahn message passing (ACMP) for graph neural networks where the numerical iteration for the particle system solution constitutes the message passing propagation. ACMP which has a simple implementation with a neural ODE solver can propel the network depth up to one hundred of layers with theoretically proven strictly positive lower bound of the Dirichlet energy. It thus provides a deep model of GNNs circumventing the common GNN problem of oversmoothing. GNNs with ACMP achieve state of the art performance for real-world node classification tasks on both homophilic and heterophilic datasets. Codes are available at https://github.com/ykiiiiii/ACMP

****

Average Sensitivity of Decision Tree Learning

Satoshi Hara,Yuichi Yoshida

A decision tree is a fundamental model used in data mining and machine learning. In practice, the training data used to construct a decision tree may change over time or contain noise, and a drastic change in the learned tree structure owing to such data perturbation is unfavorable. For example, in data mining, a change in the tree implies a change in the extracted knowledge, which raises the question of whether the extracted knowledge is truly reliable or is only a noisy artifact. To alleviate this issue, we design decision tree learning algorithms that are stable against insignificant perturbations in the training data. Specifically, we adopt the notion of average sensitivity as a stability measure, and design an algorithm with low average sensitivity that outputs a decision tree whose accuracy is close to the optimal decision tree. The experimental results on real-world datasets demonstrate that the proposed algorithm enables users to select suitable decision trees considering the trade-off between average sensitivity and accuracy.

****

Minimum Curvature Manifold Learning

Yonghyeon Lee,Frank C. Park

It is widely observed that vanilla autoencoders can have low manifold learning accuracy given a noisy or small training dataset.
Recent work has discovered that it is important to regularize the decoder that explicitly parameterizes the manifold,
where a neighborhood graph is employed for decoder regularization. However, one caveat of this method is that it is not always straightforward to construct a correct graph. Alternatively, one may consider naive graph-free regularization methods such as minimizing the norm of the decoder's Jacobian or Hessian, but these norms are not coordinate-invariant (i.e. reparametrization-invariant) and hence

do not capture any meaningful geometric quantity of the manifold nor result in geometrically meaningful manifold regularization effects.

Another recent work called the isometric regularization implicitly forces the manifold to have zero intrinsic curvature, resulting in some geometrically meaningful regularization effects. But, since the intrinsic curvature does not capture how the manifold is embedded in the data space from an extrinsic perspective, the regularization effects are often limited. In this paper, we propose a {\it minimum extrinsic curvature principle} for manifold regularization and {\bf Minimum Curvature Autoencoder (MCAE)}, a graph-free coordinate-invariant extrinsic curvature minimization framework for autoencoder regularization. Experiments with various standard datasets demonstrate that MCAE improves manifold learning accuracy compared to existing methods, especially showing strong robustness to noise.

****

Causal Proxy Models For Concept-Based Model Explanations

Zhengxuan Wu,Karel D'Oosterlinck,Atticus Geiger,Amir Zur,Christopher Potts

Explainability methods for NLP systems encounter a version of the fundamental problem of causal inference: for a given ground-truth input text, we never truly observe the counterfactual texts necessary for isolating the causal effects of model representations on outputs. In response, many explainability methods make no use of counterfactual texts, assuming they will be unavailable. In this paper, we show that robust causal explainability methods can be created using approximate counterfactuals, which can be written by humans to approximate a specific counterfactual or simply sampled using metadata-guided heuristics. The core of our proposal is the Causal Proxy Model (CPM). A CPM explains a black-box model $\mathcal{N}$ because it is trained to have the same \emph{actual} input/output behavior as $\mathcal{N}$ while creating neural representations that can be intervened upon to simulate the \emph{counterfactual} input/output behavior of $\mathcal{N}$. Furthermore, we show that the best CPM for $\mathcal{N}$ performs comparably to $\mathcal{N}$ in making factual predictions, which means that the CPM can simply replace $\mathcal{N}$, leading to more explainable deployed models.

****

Offline Reinforcement Learning with Differential Privacy

Dan Qiao,Yu-Xiang Wang

The offline reinforcement learning (RL) problem is often motivated by the need to learn data-driven decision policies in financial, legal and healthcare applications. However, the learned policy could retain sensitive information of individuals in the training data (e.g., treatment and outcome of patients), thus susceptible to various privacy risks. We design offline RL algorithms with differential privacy guarantees which provably prevent such risks. These algorithms also enjoy strong instance-dependent learning bounds under both tabular and linear Markov Decision Process (MDP) settings. Our theory and simulation suggest that the privacy guarantee comes at (almost) no drop in utility comparing to the non-private counterpart for a medium-size dataset.

****

Relational Attention: Generalizing Transformers for Graph-Structured Tasks

Cameron Diao,Ricky Loynd

Transformers flexibly operate over sets of real-valued vectors representing task-specific entities and their attributes, where each vector might encode one word-piece token and its position in a sequence, or some piece of information that carries no position at all. As set processors, transformers are at a disadvantage in reasoning over more general graph-structured data where nodes represent entities and edges represent relations between entities. To address this shortcoming, we generalize transformer attention to consider and update edge vectors in each transformer layer. We evaluate this relational transformer on a diverse array of graph-structured tasks, including the large and challenging CLRS Algorithmic Reasoning Benchmark. There, it dramatically outperforms state-of-the-art graph neural networks expressly designed to reason over graph-structured data. Our analysis demonstrates that these gains are attributable to relational attention's inherent ability to leverage the greater expressivity of graphs over sets.

****

Distilling Model Failures as Directions in Latent Space
Saachi Jain,Hannah Lawrence,Ankur Moitra,Aleksander Madry

Existing methods for isolating hard subpopulations and spurious correlations in datasets often require human intervention. This can make these methods labor-intensive and dataset-specific. To address these shortcomings, we present a scalable method for automatically distilling a model's failure modes. Specifically, we harness linear classifiers to identify consistent error patterns, and, in turn, induce a natural representation of these failure modes as directions within the feature space. We demonstrate that this framework allows us to discover and automatically caption challenging subpopulations within the training dataset. Moreover, by combining our framework with off-the-shelf diffusion models, we can generate images that are especially challenging for the analyzed model, and thus can be used to perform synthetic data augmentation that helps remedy the model's failure modes.

****

Stable Target Field for Reduced Variance Score Estimation in Diffusion Models
Yilun Xu,Shangyuan Tong,Tommi S. Jaakkola

Diffusion models generate samples by reversing a fixed forward diffusion process. Despite already providing impressive empirical results, these diffusion models algorithms can be further improved by reducing the variance of the training targets in their denoising score-matching objective. We argue that the source of such variance lies in the handling of intermediate noise-variance scales, where multiple modes in the data affect the direction of reverse paths. We propose to remedy the problem by incorporating a reference batch which we use to calculate weighted conditional scores as more stable training targets. We show that the procedure indeed helps in the challenging intermediate regime by reducing (the trace of) the covariance of training targets. The new stable targets can be seen as trading bias for reduced variance, where the bias vanishes with increasing reference batch size. Empirically, we show that the new objective improves the image quality, stability, and training speed of various popular diffusion models across datasets with both general ODE and SDE solvers. When used in combination with EDM, our method yields a current SOTA FID of 1.90 with 35 network evaluations on the unconditional CIFAR-10 generation task. The code is available at https://github.com/Newbeeer/stf

****

Graph Contrastive Learning Under Heterophily: Utilizing Graph Filters to Generate Graph Views
Wenhan Yang,Baharan Mirzasoleiman

Graph Neural Networks have achieved tremendous success in (semi-)supervised tasks for which task-specific node labels are available. However, obtaining labels is expensive in many domains, specially as the graphs grow larger in size. Hence, there has been a growing interest in the application of self-supervised techniques, in particular contrastive learning (CL), to graph data. In general, CL methods work by maximizing the agreement between encoded augmentations of the same example, and minimizing agreement between encoded augmentations of different examples. However, we show that existing graph CL methods perform very poorly on graphs with heterophily, in which connected nodes tend to belong to different classes. First, we show that this is attributed to the ineffectiveness of existing graph augmentation methods. Then, we leverage graph filters to directly generate augmented graph views for graph CL under heterophily. In particular, instead of explicitly augmenting the graph topology and encoding the augmentations, we use a high-pass filter in the encoder to generate node representations only based on high-frequency graph signals. Then, we contrast the high-pass filtered representations with their low-pass counterparts produced by the same encoder, to generate representations. Our experimental results confirm that our proposed method, HLCL, outperforms state-of-the-art CL methods on benchmark graphs with heterophily, by up to 10%.

****

Countinuous pseudo-labeling from the start
Dan Berrebbi,Ronan Collobert,Samy Bengio,Navdeep Jaitly,Tatiana Likhomanenko

Self-training (ST), or pseudo-labeling has sparked significant interest in the automatic speech recognition (ASR) community recently because of its success in harnessing unlabeled data. Unlike prior semi-supervised learning approaches that relied on iteratively regenerating pseudo-labels (PLs) from a trained model and using them to train a new model, recent state-of-the-art methods perform `continuous training' where PLs are generated using a very recent version of the model being trained. Nevertheless, these approaches still rely on bootstrapping the ST using an initial supervised learning phase where the model is trained on labeled data alone. We believe this has the potential for over-fitting to the labeled dataset in low resource settings and that ST from the start of training should reduce over-fitting. In this paper we show how we can do this by dynamically controlling the evolution of PLs during the training process in ASR. To the best of our knowledge, this is the first study that shows the feasibility of generating PLs from the very start of the training. We are able to achieve this using two techniques that avoid instabilities which lead to degenerate models that do not generalize. Firstly, we control the evolution of PLs through a curriculum that uses the online changes in PLs to control the membership of the cache of PLs and improve generalization. Secondly, we find that by sampling transcriptions from the predictive distribution, rather than only using the best transcription, we can stabilize training further. With these techniques, our ST models match prior works without an external language model.

****

Hybrid Federated Learning for Feature & Sample Heterogeneity: Algorithms and Implementation

Xinwei Zhang,Wotao Yin,Mingyi Hong,Tianyi Chen

Federated learning (FL) is a popular distributed machine learning paradigm dealing with distributed and private data sets. Based on the data partition pattern, FL is often categorized into horizontal, vertical, and hybrid settings. All three settings have many applications, but the hybrid FL remains relatively less explored, because it deals with the challenging situation where both the feature space and the data samples are heterogeneous.

This work designs a novel mathematical model that effectively allows the clients to aggregate distributed data with heterogeneous, and possibly overlapping features and samples. Our main idea is to partition each client's model into a feature extractor part and a classifier part, where the former can be used to process the input data, while the latter is used to perform the learning from the extracted features. The heterogeneous feature aggregation is done through building a server model, which assimilates local classifiers and feature extractors through a carefully designed matching mechanism. A communication-efficient algorithm is then designed to train both the client and server models. Finally, we conducted numerical experiments on multiple image classification data sets to validate the performance of the proposed algorithm. To our knowledge, this is the first formulation and algorithm developed for hybrid FL.

****

Policy Architectures for Compositional Generalization in Control

Allan Zhou,Vikash Kumar,Chelsea Finn,Aravind Rajeswaran

Several tasks in control, robotics, and planning can be specified through desired goal configurations for entities in the environment. Learning goal-conditioned policies is a natural paradigm to solve such tasks. However, learning and generalizing on complex tasks can be challenging due to variations in number of entities or compositions of goals. To address this challenge, we introduce the Entity-Factored Markov Decision Process (EFMDP), a formal framework for modeling the entity-based compositional structure in control tasks. Geometrical properties of the EFMDP framework provide theoretical motivation for policy architecture design, particularly Deep Sets and popular relational mechanisms such as graphs and self attention. These structured policy architectures are flexible and can be trained end-to-end with standard reinforcement and imitation learning algorithms. We study and compare the learning and generalization properties of these architectures on a suite of simulated robot manipulation tasks, finding that they achieve significantly higher success rates with less data compared to standard multila

yer perceptrons. Structured policies also enable broader and more compositional generalization, producing policies that extrapolate to different numbers of entities than seen in training, and stitch together (i.e. compose) learned skills in novel ways. Video results can be found at https://sites.google.com/view/comp-gen-anon.
****
GNNDelete: A General Strategy for Unlearning in Graph Neural Networks
Jiali Cheng,George Dasoulas,Huan He,Chirag Agarwal,Marinka Zitnik
Graph unlearning, which involves deleting graph elements such as nodes, node labels, and relationships from a trained graph neural network (GNN) model, is crucial for real-world applications where data elements may become irrelevant, inaccurate, or privacy-sensitive. However, existing methods for graph unlearning either deteriorate model weights shared across all nodes or fail to effectively delete edges due to their strong dependence on local graph neighborhoods. To address these limitations, we introduce GNNDelete, a novel model-agnostic layer-wise operator that optimizes two critical properties, namely, Deleted Edge Consistency and Neighborhood Influence, for graph unlearning. Deleted Edge Consistency ensures that the influence of deleted elements is removed from both model weights and neighboring representations, while Neighborhood Influence guarantees that the remaining model knowledge is preserved after deletion. GNNDelete updates representations to delete nodes and edges from the model while retaining the rest of the learned knowledge. We conduct experiments on seven real-world graphs, showing that GNNDelete outperforms existing approaches by up to 38.8% (AUC) on edge, node, and node feature deletion tasks, and 32.2% on distinguishing deleted edges from non-deleted ones. Additionally, GNNDelete is efficient, taking 12.3x less time and 9.3x less space than retraining GNN from scratch on WordNet18.
****
Lower Bounds for Differentially Private ERM: Unconstrained and Non-Euclidean
Zhou Lu,Daogao Liu
We consider the lower bounds of differentially private empirical risk minimization (DP-ERM) for convex functions in both constrained and unconstrained cases concerning the general ℓ_p norm beyond the ℓ_2 norm considered by most of the previous works.

We provide a simple black-box reduction approach that can generalize lower bounds in constrained to unconstrained cases.
Moreover, for (ϵ,δ)-DP, we achieve the optimal $\Omega(\frac{\sqrt{d \log(1/\delta)}}{\epsilon n})$ lower bounds for both constrained and unconstrained cases and any ℓ_p geometry where $p\geq 1$ by considering ℓ_1 loss over the ℓ_{∞} ball.
****
The Convergence Rate of SGD's Final Iterate: Analysis on Dimension Dependence
Daogao Liu,Zhou Lu
Stochastic Gradient Descent (SGD) is among the simplest and most popular optimization and machine learning methods. Running SGD with a fixed step size and outputting the final iteration is an ideal strategy one can hope for, but it is still not well-understood even though SGD has been studied extensively for over 70 years. Given the $\Theta(\log T)$ gap between current upper and lower bounds for running SGD for T steps, it was then asked by [Koren and Segal COLT 20'] how to characterize the final-iterate convergence of SGD with a fixed step size in the constant dimension setting, i.e., $d=O(1)$.

In this paper, we consider the more general setting for any $d\leq T$, proving $\Omega(\log d/\sqrt{T})$ lower bounds for the sub-optimality of the final iterate of SGD in minimizing non-smooth Lipschitz convex functions with standard step sizes. Our results provide the first general dimension-dependent lower bound on the convergence of SGD's final iterate, partially resolving the COLT open question raised by [Koren and Segal COLT 20'].
Moreover, we present a new method in one dimension based on martingale and Freedman’s inequality, which gets the tight $O(1/\sqrt{T})$ upper bound with mild ass

umptions and recovers the same bounds $O(\log T/\sqrt{T})$ as previous best results under the standard assumptions.

****

Multi-Rate VAE: Train Once, Get the Full Rate-Distortion Curve

Juhan Bae,Michael R. Zhang,Michael Ruan,Eric Wang,So Hasegawa,Jimmy Ba,Roger Baker Grosse

Variational autoencoders (VAEs) are powerful tools for learning latent representations of data used in a wide range of applications. In practice, VAEs usually require multiple training rounds to choose the amount of information the latent variable should retain. This trade-off between the reconstruction error (distortion) and the KL divergence (rate) is typically parameterized by a hyperparameter β. In this paper, we introduce Multi-Rate VAE (MR-VAE), a computationally efficient framework for learning optimal parameters corresponding to various β in a single training run. The key idea is to explicitly formulate a response function using hypernetworks that maps β to the optimal parameters. MR-VAEs construct a compact response hypernetwork where the pre-activations are conditionally gated based on β. We justify the proposed architecture by analyzing linear VAEs and showing that it can represent response functions exactly for linear VAEs. With the learned hypernetwork, MR-VAEs can construct the rate-distortion curve without additional training and can be deployed with significantly less hyperparameter tuning. Empirically, our approach is competitive and often exceeds the performance of multiple β-VAEs training with minimal computation and memory overheads.

****

Adaptive Gradient Methods with Local Guarantees

Zhou Lu,Wenhan Xia,Sanjeev Arora,Elad Hazan

Adaptive gradient methods are the method of choice for optimization in machine learning and used to train the largest deep models. In this paper we study the problem of learning a local preconditioner, that can change as the data is changing along the optimization trajectory. We propose an adaptive gradient method that has provable adaptive regret guarantees vs. the best local preconditioner. To derive this guarantee, we prove a new adaptive regret bound in online learning that improves upon previous adaptive online learning methods.

We demonstrate the robustness of our method in automatically choosing the optimal learning rate schedule for popular benchmarking tasks in vision and language domains. Without the need to manually tune a learning rate schedule, our method can, in a single run, achieve comparable and stable task accuracy as a fine-tuned optimizer.

****

Combinatorial-Probabilistic Trade-Off: P-Values of Community Properties Test in the Stochastic Block Models

Shuting Shen,Junwei Lu

We propose an inferential framework testing the general community combinatorial properties of the stochastic block model. We aim to test the hypothesis on whether a certain community property is satisfied, e.g., whether a given set of nodes belong to the same community, and provide p-values for uncertainty quantification. Our framework is applicable to all symmetric community properties. To ease the challenges caused by the combinatorial nature of community properties, we develop a novel shadowing bootstrap method. Utilizing the symmetry, our method can find a shadowing representative of the true assignment and the number of tested assignments in the alternative is largely reduced. In theory, we introduce a combinatorial distance between two community classes and show a combinatorial-probabilistic trade-off phenomenon. Our test is honest as long as the product of the combinatorial distance between two communities and the probabilistic distance between two connection probabilities is sufficiently large. Besides, we show that such trade-off also exists in the information-theoretic lower bound. We also implement numerical experiments to show the validity of our method.

****

Min-Max Zero-Shot Multi-Label Classification

Sima Behpour

In many classification problems, acquiring labeled examples for many classes is difficult, resulting in high interest in zero-shot learning frameworks. Zero-shot learning (ZSL) is a problem setup, where at test time, a learner observes samples from classes that were not observed/trained in the training phase and is required to predict the category they belong to. Zero-shot learning transfers knowledge from seen classes (observed classes in the training phase) to unseen classes (unobserved classes in the training phase but present in the testing phase), reducing human labor of data annotation to build new classifiers. However, most zero-shot learning researches target single-label classification (multi-class setting). There are few studies on multi-label zero-shot learning due to the difficulty in modeling complex semantics conveyed by a set of labels.

We propose a novel probabilistic model that incorporates more general feature representation (e.g., Word-Net hierarchy, word2vec features, convolutional neural network features (layer-wise), and co-occurrence statistics) and learns the knowledge transfer in terms of data structure and relations. We also investigate the effect of leveraging different CNN layers' features. Our experimental results prove the efficacy of our model in handling unseen labels. We run additional experiments to analyze the flat-sharp minima convergence of methods as a generalization factor. Our study suggests that our proposed method converges to flat minima resulting in strong generalization.

****

Meta-Learning in Games

Keegan Harris,Ioannis Anagnostides,Gabriele Farina,Mikhail Khodak,Steven Wu,Tuomas Sandholm

In the literature on game-theoretic equilibrium finding, focus has mainly been on solving a single game in isolation. In practice, however, strategic interactions—ranging from routing problems to online advertising auctions—evolve dynamically, thereby leading to many similar games to be solved. To address this gap, we introduce meta-learning for equilibrium finding and learning to play games. We establish the first meta-learning guarantees for a variety of fundamental and well-studied games, including two-player zero-sum games, general-sum games, Stackelberg games, and multiple extensions thereof. In particular, we obtain rates of convergence to different game-theoretic equilibria that depend on natural notions of similarity between the sequence of games encountered, while at the same time recovering the known single-game guarantees when the sequence of games is arbitrary. Along the way, we prove a number of new results in the single-game regime through a simple and unified framework, which may be of independent interest. Finally, we evaluate our meta-learning algorithms on endgames faced by the poker agent Libratus against top human professionals. The experiments show that games with varying stack sizes can be solved significantly faster using our meta-learning techniques than by solving them separately, often by an order of magnitude.

****

GLASU: A Communication-Efficient Algorithm for Federated Learning with Vertically Distributed Graph Data

Xinwei Zhang,Mingyi Hong,Jie Chen

Vertical federated learning (VFL) is a distributed learning paradigm, where computing clients collectively train a model based on the partial features of the same set of samples they possess. Current research on VFL focuses on the case when samples are independent, but it rarely addresses an emerging scenario when samples are interrelated through a graph. For graph-structured data, graph neural networks (GNNs) are rather competitive machine learning models, but a naive implementation in the VFL setting causes a significant communication overhead; moreover, the analysis is faced with a challenge caused by the biased stochastic gradients. In this paper, we propose a model splitting method that splits a backbone GNN across the clients and the server and a communication-efficient algorithm, GLASU, to train such a model. GLASU adopts lazy aggregation and stale updates to skip aggregation when evaluating the model and skip feature exchanges during training, greatly reducing communication. We offer a theoretical analysis and conduc t extensive numerical experiments on real-world datasets, showing that the proposed algorithm effectively trains a GNN model, whose performance matches that of the backbone GNN when trained in a centralized manner.

****

Dynamic Embeddings of Temporal High-Order Interactions via Neural Diffusion-Reaction Processes

Zheng Wang,shikai fang,Shibo Li,Shandian Zhe

High-order interactions of multiple entities are ubiquitous in practical applications. The associated data often includes the participants, interaction results, and the timestamps when each interaction occurred. While tensor factorization is a popular tool to analyze such data, it often ignores or underuses valuable timestamp information. More important, standard tensor factorization only estimates a static representation for each entity and ignores the temporal variation of the representations. However, such variations might reflect important evolution patterns of the underlying properties of the entities. To address these limitations, we propose Dynamical eMbedIngs of TempoRal hIgh-order interactions (DMITRI). We develop a neural diffusion-reaction process model to estimate the dynamic embeddings for the participant entities. Specifically, based on the observed interactions, we build a multi-partite graph to encode the correlation between the entities. We construct a graph diffusion process to co-evolve the embedding trajectories of the correlated entities and use a neural network to construct a reaction process for each individual entity. In this way, our model is able to capture both the commonalities and personalities during the evolution of the embeddings for different entities. We then use a neural network to model the interaction result as a nonlinear function of the embedding trajectories. For model estimation, we combine ODE solvers to develop a stochastic mini-batch learning algorithm. We propose a simple stratified sampling method to balance the cost of processing each mini-batch so as to improve the overall efficiency. We show the advantage of our approach in both the ablation study and real-world applications.

****

Fair Federated Learning via Bounded Group Loss

Shengyuan Hu,Steven Wu,Virginia Smith

In federated learning, fair prediction across protected groups is an important constraint for many applications. Unfortunately, prior work studying group fair federated learning lacks formal convergence or fairness guarantees. In this work we propose a general framework for provably fair federated learning. In particular, we explore and extend the notion of Bounded Group Loss as a theoretically-grounded approach for group fairness. Using this setup, we propose a scalable federated optimization method that optimizes the empirical risk under a number of group fairness constraints. We provide convergence guarantees for the method as well as fairness guarantees for the resulting solution. Empirically, we evaluate our method across common benchmarks from fair ML and federated learning, showing that it can provide both fairer and more accurate predictions than baseline approaches.

****

DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking

Gabriele Corso,Hannes Stärk,Bowen Jing,Regina Barzilay,Tommi S. Jaakkola

Predicting the binding structure of a small molecule ligand to a protein---a task known as molecular docking---is critical to drug design. Recent deep learning methods that treat docking as a regression problem have decreased runtime compared to traditional search-based methods but have yet to offer substantial improvements in accuracy. We instead frame molecular docking as a generative modeling problem and develop DiffDock, a diffusion generative model over the non-Euclidean manifold of ligand poses. To do so, we map this manifold to the product space of the degrees of freedom (translational, rotational, and torsional) involved in docking and develop an efficient diffusion process on this space. Empirically, DiffDock obtains a 38% top-1 success rate (RMSD<2A) on PDBBind, significantly outperforming the previous state-of-the-art of traditional docking (23%) and deep learning (20%) methods. Moreover, while previous methods are not able to dock on computationally folded structures (maximum accuracy 10.4%), DiffDock maintains s ignificantly higher precision (21.7%). Finally, DiffDock has fast inference times and provides confidence estimates with high selective accuracy.

****

An Upper Bound for the Distribution Overlap Index and Its Applications
Hao Fu,Prashanth Krishnamurthy,Siddharth Garg,Farshad Khorrami

The overlap index between two probability distributions has various applications in statistics, machine learning, and other scientific research. However, approximating the overlap index is challenging when the probability distributions are unknown (i.e., distribution-free settings). This paper proposes an easy-to-compute upper bound for the overlap index without requiring any knowledge of the distribution models. We first utilize the bound to find the upper limit for the accuracy of a trained machine learning model when a domain shift exists. We additionally employ this bound to study the distribution membership classification of given samples. Specifically, we build a novel, distribution-free, computation-efficient, memory-efficient one-class classifier by converting the bound into a confidence score function. The proposed classifier does not need to train any parameters and requires only a small number of in-class samples to be accurate. The classifier shows its efficacy on various datasets and outperforms many state-of-the-art methods in different one-class classification scenarios, including novelty detection, out-of-distribution detection, and backdoor detection. The obtained results show significant promise toward broadening the applications of overlap-based metrics.

****

Learning by Distilling Context
Charlie Victor Snell,Dan Klein,Ruiqi Zhong

Language models significantly benefit from context tokens, such as prompts or scratchpads. They perform better when prompted with concrete training examples and abstract statements about the target task (instructions), and they acquire new capabilities to perform complex tasks by generating step-by-step reasoning (scratch-pad) before predicting the final answers. However, they do not internalize these performance gains, which disappear when the context tokens are gone. Consequently, we always need to pay extra computation for this gain, and it is unclear how to transfer the capabilities acquired by context tokens to other tasks, or how to leverage the context tokens when their length exceeds the context window size. Our work proposes to apply context distillation so that a language model can internalize these gains. Concretely, given an input for the target task, we let the model use all relevant context tokens to generate the output, using ``[instructions] + [task-input]'' to predict ``[scratch-pad] + [final answer]''; then we fine-tune the same model to predict the above ``[final answer]'' conditioned on the ``[task-input]'', without seeing the ``[instructions]'' or using the ``[scratch-pad]''. This incentivizes the model to behave as if the context were present, hence updating the parameters to internalize the context information. We show that context distillation can be used as a general method for learning. In particular, we demonstrate that context distillation can effectively internalize 3 types of contexts: 1) abstract task instructions and natural language explanations of why an output is correct or incorrect on Natural-Instructions-V2; 2) step-by-step reasoning on 8-digit addition questions, where we show the model can apply this newly acquired capability to downstream question answering tasks; and 3) concrete training examples on the SPIDER Text-to-SQL dataset, where context distillation outperforms directly learning with gradient descent by 7%.

****

Target-Free Ligand Scoring via One-Shot Learning
Peter Eckmann,Jake Anderson,Michael K Gilson,Rose Yu

Scoring ligands in a library based on their structural similarity to a known hit compound is widely used in drug discovery following high-throughput screening. However, such "similarity search" relies on the assumption that structurally similar compounds have similar activities, and will therefore only retrieve ligands with hit-like affinity, requiring resource-intensive tweaking by medicinal chemists to reach a more active lead compound. We propose a novel approach, One-Shot Ligand Scoring (OSLS), that is much more capable of directly retrieving lead-li ke compounds from a library using a novel one-shot learning technique. For this new task, we design a Siamese-inspired neural architecture using two Transformer encoders without tied weights, a novel positional encoding-like mechanism, and a final prediction head. OSLS is able to score ligands by activity against a target without any target-specific knowledge beyond a single known activity value, a cost-effective approach to ligand-based or phenotypic drug discovery. We show that OSLS surpasses traditional similarity search as well as modern deep learning baselines on a simulated ligand retrieval task. Furthermore, we demonstrate the applicability of our approach on various drug discovery tasks that also involve ligand scoring, including drug repositioning, precision patient-level drug efficacy prediction, and even molecular generative modeling.

****

Structured Pruning of CNNs at Initialization

Yaohui Cai,Weizhe Hua,Hongzheng Chen,Fengyu Li,G. Edward Suh,Christopher De Sa,Zhiru Zhang

Pruning-at-initialization (PAI) proposes to prune the individual weights of the CNN before training, thus avoiding expensive fine-tuning or retraining of the pruned model. While PAI shows promising results in reducing model size, the pruned model still requires unstructured sparse matrix computation, making it difficult to achieve wall-clock speedups. In this work, we show theoretically and empirically that the accuracy of CNN models pruned by PAI methods only depends on the fraction of remaining parameters in each layer (i.e., layer-wise density), regardless of the granularity of pruning. We formulate the PAI problem as a convex optimization of our newly proposed expectation-based proxy for model accuracy, which leads to finding the optimal layer-wise density of that specific model. Based on our formulation, we further propose a structured and hardware-friendly PAI method, named PreCrop, to prune or reconfigure CNNs in the channel dimension. Our empirical results show that PreCrop achieves a higher accuracy than existing PAI methods on several modern CNN architectures, including ResNet, MobileNetV2, and EfficientNet for both CIFAR-10 and ImageNet. PreCrop achieves an accuracy improvement of up to 2.7% over the state-of-the-art PAI algorithm when pruning MobileNetV2 on ImageNet. PreCrop also improves the accuracy of EfficientNetB0 by 0.3% on ImageNet with only 80% of the parameters and the same FLOPs.

****

Constructive TT-representation of the tensors given as index interaction functions with applications

Gleb Ryzhakov,Ivan Oseledets

This paper presents a method to build explicit tensor-train (TT) representations. We show that a wide class of tensors can be explicitly represented with sparse TT-cores, obtaining, in many cases, optimal TT-ranks. Numerical experiments show that our method outperforms the existing ones in several practical applications, including game theory problems. Theoretical estimations of the number of operations show that in some problems, such as permanent calculation, our methods are close to the known optimal asymptotics, which are obtained by a completely different type of methods.

****

Thinking Two Moves Ahead: Anticipating Other Users Improves Backdoor Attacks in Federated Learning

Yuxin Wen,Jonas Geiping,Liam H Fowl,Hossein Souri,Rama Chellappa,Micah Goldblum,Tom Goldstein

Federated learning is particularly susceptible to model poisoning and backdoor attacks because individual users have direct control over the training data and model updates. At the same time, the attack power of an individual user is limited because their updates are quickly drowned out by those of many other users. Existing attacks do not account for future behaviors of other users, and thus require many sequential updates and their effects are quickly erased. We propose an attack that anticipates and accounts for the entire federated learning pipeline, including behaviors of other clients, and ensures that backdoors are effective quickly and persist even after multiple rounds of community updates. We show th at this new attack is effective in realistic scenarios where the attacker only contributes to a small fraction of randomly sampled rounds and demonstrate this attack on image classification, next-word prediction, and sentiment analysis.

****

Continuized Acceleration for Quasar Convex Functions in Non-Convex Optimization

Jun-Kun Wang,Andre Wibisono

Quasar convexity is a condition that allows some first-order methods to efficiently minimize a function even when the optimization landscape is non-convex. Previous works develop near-optimal accelerated algorithms for minimizing this class of functions, however, they require a subroutine of binary search which results in multiple calls to gradient evaluations in each iteration, and consequently the total number of gradient evaluations does not match a known lower bound. In this work, we show that a recently proposed continuized Nesterov acceleration can be applied to minimizing quasar convex functions and achieves the optimal bound with a high probability. Furthermore, we find that the objective functions of training generalized linear models (GLMs) satisfy quasar convexity, which broadens the applicability of the relevant algorithms, while known practical examples of quasar convexity in non-convex learning are sparse in the literature. We also show that if a smooth and one-point strongly convex, Polyak-Lojasiewicz, or quadratic-growth function satisfies quasar convexity, then attaining an accelerated linear rate for minimizing the function is possible under certain conditions, while acceleration is not known in general for these classes of functions.

****

Towards Global Optimality in Cooperative MARL with Sequential Transformation

Jianing Ye,Chenghao Li,Jianhao Wang,Qianchuan Zhao,Chongjie Zhang

Policy learning in multi-agent reinforcement learning (MARL) is challenging due to the exponential growth of joint state-action space with respect to the number of agents. To achieve higher scalability, the paradigm of centralized training with decentralized execution (CTDE) is broadly adopted with factorized structure in MARL. However, we observe that existing CTDE algorithms in cooperative MARL cannot achieve optimality even in simple matrix games. To understand this phenomenon, we analyze two mainstream classes of CTDE algorithms -- actor-critic algorithms and value-decomposition algorithms. Our theoretical and experimental results characterize the weakness of these two classes of algorithms when the optimization method is taken into consideration, which indicates that the currently used centralized training manner is deficient in compatibility with decentralized policy. To address this issue, we present a transformation framework that reformulates a multi-agent MDP as a special "single-agent" MDP with a sequential structure and can allow employing off-the-shelf single-agent reinforcement learning (SARL) algorithms to efficiently learn corresponding multi-agent tasks. After that, a decentralized policy can still be learned by distilling the "single-agent" policy. This framework retains the optimality guarantee of SARL algorithms into cooperative MARL. To instantiate this transformation framework, we propose a Transformed PPO, called T-PPO, which can theoretically perform optimal policy learning in the finite multi-agent MDPs and shows significant outperformance on a large set of cooperative multi-agent tasks.

****

Sparse tree-based Initialization for Neural Networks

Patrick Lutz,Ludovic Arnould,Claire Boyer,Erwan Scornet

Dedicated neural network (NN) architectures have been designed to handle specific data types (such as CNN for images or RNN for text), which ranks them among state-of-the-art methods for dealing with these data. Unfortunately, no architecture has been found for dealing with tabular data yet, for which tree ensemble methods (tree boosting, random forests) usually show the best predictive performances. In this work, we propose a new sparse initialization technique for (potentially deep) multilayer perceptrons (MLP): we first train a tree-based procedure to detect feature interactions and use the resulting information to initialize the network, which is subsequently trained via standard gradient descent (GD) strategies. Numerical experiments on several tabular data sets showthe benefits of t his new, simple and easy-to-use method, both in terms of generalization capacity and computation time, compared to default MLP initialization and even to existing complex deep learning solutions. In fact, this wise MLP initialization raises the performances of the resulting NN methods to that of gradient boosting on tabular data. Besides, such initializations are able to preserve the sparsity of weights introduced in the first layers of the network throughout the training, which emphasizes that the first layers act as a sparse feature extractor (like convolutional layers in CNN).

****

Learning Soft Constraints From Constrained Expert Demonstrations

Ashish Gaurav,Kasra Rezaee,Guiliang Liu,Pascal Poupart

Inverse reinforcement learning (IRL) methods assume that the expert data is generated by an agent optimizing some reward function. However, in many settings, the agent may optimize a reward function subject to some constraints, where the constraints induce behaviors that may be otherwise difficult to express with just a reward function. We consider the setting where the reward function is given, and the constraints are unknown, and propose a method that is able to recover these constraints satisfactorily from the expert data. While previous work has focused on recovering hard constraints, our method can recover cumulative soft constraints that the agent satisfies on average per episode. In IRL fashion, our method solves this problem by adjusting the constraint function iteratively through a constrained optimization procedure, until the agent behavior matches the expert behavior. We demonstrate our approach on synthetic environments, robotics environments and real world highway driving scenarios.

****

VoGE: A Differentiable Volume Renderer using Gaussian Ellipsoids for Analysis-by-Synthesis

Angtian Wang,Peng Wang,Jian Sun,Adam Kortylewski,Alan Yuille

Differentiable rendering allows the application of computer graphics on vision tasks, e.g. object pose and shape fitting, via analysis-by-synthesis, where gradients at occluded regions are important when inverting the rendering process.To obtain those gradients, state-of-the-art (SoTA) differentiable renderers use rasterization to collect a set of nearest components for each pixel and aggregate them based on the viewing distance. In this paper, we propose VoGE, which uses ray tracing to capture nearest components with their volume density distributions on the rays and aggregates via integral of the volume densities based on Gaussian ellipsoids, which brings more efficient and stable gradients. To efficiently render via VoGE, we propose an approximate close-form solution for the volume density aggregation and a coarse-to-fine rendering strategy. Finally, we provide a CUDA implementation of VoGE, which gives a competitive rendering speed in comparison to PyTorch3D. Quantitative and qualitative experiment results show VoGE outperforms SoTA counterparts when applied to various vision tasks, e.g., object pose estimation, shape/texture fitting, and occlusion reasoning. The VoGE code is available at: https://github.com/Angtian/VoGE.

****

Unravel Structured Heterogeneity of Tasks in Meta-Reinforcement Learning via Exploratory Clustering

Zhendong Chu,Hongning Wang

Meta-reinforcement learning (meta-RL) is developed to quickly solve new tasks by leveraging knowledge from prior tasks. The assumption that tasks are drawn IID is typically made in previous studies, which ignore possible structured heterogeneity of tasks. The non-transferable knowledge caused by structured heterogeneity hinders fast adaptation in new tasks. In this paper, we formulate the structured heterogeneity of tasks via clustering such that transferable knowledge can be inferred within different clusters and non-transferable knowledge would be excluded across clusters thereby. To facilitate so, we develop a dedicated exploratory policy to discover task clusters by reducing uncertainty in posterior inference. Within the identified clusters, the exploitation policy is able to solve related tasks by utilizing knowledge shared within the clusters. Experiments on various MuJoCo tasks showed the proposed method can unravel cluster structures effe ctively in both rewards and state dynamics, proving strong advantages against a set of state-of-the-art baselines.

****

An Investigation of Domain Generalization with Rademacher Complexity

Da Li,Henry Gouk,Timothy Hospedales

The domain generalization (DG) setting challenges a model trained on multiple known data distributions to generalise well on unseen data distributions. Due to its practical importance, many methods have been proposed to address this challenge.

However much work in general purpose DG is heuristically motivated,
as the DG problem is hard to model formally; and recent evaluations have cast doubt on existing methods’ practical efficacy -- in particular compared to a well tuned empirical risk minimisation baseline.

We present a novel learning-theoretic generalisation bound for DG that bounds unseen domain generalisation performance in terms of the model’s empirical risk and Rademacher complexity -- providing a sufficient condition for DG. Based on this insight, we empirically analyze the performance of several methods and show that their performance is indeed influenced by model complexity in practice.

Algorithmically, our analysis suggests that tuning for domain generalisation should be achieved by simply performing regularised ERM with a leave-one-domain-out cross-validation objective. Empirical results on the DomainBed benchmark corroborate this.

****

FedDA: Faster Framework of Local Adaptive Gradient Methods via Restarted Dual Averaging

Junyi Li,Feihu Huang,Heng Huang

Federated learning (FL) is an emerging learning paradigm to tackle massively distributed data. In Federated Learning, a set of clients jointly perform a machine learning task under the coordination of a server. The FedAvg algorithm is one of the most widely used methods to solve Federated Learning problems. In FedAvg, the learning rate is a constant rather than changing adaptively. The adaptive gradient methods show superior performance over the constant learning rate schedule; however, there is still no general framework to incorporate adaptive gradient methods into the federated setting. In this paper, we propose \textbf{FedDA}, a novel framework for local adaptive gradient methods. The framework adopts a restarted dual averaging technique and is flexible with various gradient estimation methods and adaptive learning rate formulations. In particular, we analyze \textbf{FedDA-MVR}, an instantiation of our framework, and show that it achieves gradient complexity $\tilde{O}(\epsilon^{-1.5})$ and communication complexity $\tilde{O}(\epsilon^{-1})$ for finding a stationary point ϵ. This matches the best known rate for first-order FL algorithms and \textbf{FedDA-MVR} is the first adaptive FL algorithm that achieves this rate. We also perform extensive numerical experiments to verify the efficacy of our method.

****

The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers

Zonglin Li,Chong You,Srinadh Bhojanapalli,Daliang Li,Ankit Singh Rawat,Sashank J. Reddi,Ke Ye,Felix Chern,Felix Yu,Ruiqi Guo,Sanjiv Kumar

This paper studies a curious phenomenon that machine learning model with Transformer architectures have sparse activation maps. By activation map we refer to the intermediate output of the multi-layer perceptrons (MLPs) after a ReLU activation function, and by "sparse" we mean that on average very few entries (e.g., 3.0% for T5-Base and 6.3% for ViT-B16) are nonzero for each input to MLP. Moreover, larger Transformers with more layers and wider MLP hidden dimensions are sparser as measured by the percentage of nonzero entries. Through extensive experiments we demonstrate that the emergence of sparsity is a prevalent phenomenon that occurs for both natural language processing and vision tasks, on both training and evaluation data, for Transformers of various configurations, at layers of all depth levels. We discuss how sparsity immediately implies a way to significantly reduce the FLOP count and improve efficiency for Transformers. Moreover, we demonstrate perhaps surprisingly that enforcing an even sparser activation via Top

-k thresholding with a small k brings a collection of desired properties, namely less sensitivity to noisy training data, more robustness to input corruptions, and better calibration for their prediction confidence.

****

Explainable Recommender with Geometric Information Bottleneck

Hanqi Yan,Lin Gui,Yulan He

Explainable recommender systems have attracted much interest in recent years as they can explain their recommendation decisions, enhancing user trust in the systems. Most explainable recommender systems rely on human-generated rationales or annotated aspect features from user reviews to train models for rational generation or extraction. The rationales produced are often confined to a single review. To avoid the expensive human annotation process and to generate explanations beyond individual reviews, we propose an explainable recommender system trained on user reviews by developing a transferable Geometric Information Bottleneck (GIANT), which leverages the prior knowledge acquired through clustering on a user-item graph built on user-item rating interactions, since graph nodes in the same cluster tend to share common characteristics or preferences. We then feed user reviews and item reviews into a variational network to learn latent topic distributions which are regularised by the distributions of user/item estimated based on their distances to various cluster centroids of the user-item graph. By iteratively refining the instance-level review latent topics with GIANT, our method learns a robust latent space from the text for rating prediction and explanation generation. Experimental results on three e-commerce datasets show that our model significantly improves the interpretability of a variational recommender using a standard Gaussian prior, in terms of coherence, diversity and faithfulness, while achieving performance comparable to existing content-based recommender systems in terms of rating prediction accuracy.

****

Near-optimal Policy Identification in Active Reinforcement Learning

Xiang Li,Viraj Mehta,Johannes Kirschner,Ian Char,Willie Neiswanger,Jeff Schneider,Andreas Krause,Ilija Bogunovic

Many real-world reinforcement learning tasks require control of complex dynamical systems that involve both costly data acquisition processes and large state spaces. In cases where the expensive transition dynamics can be readily evaluated at specified states (e.g., via a simulator), agents can operate in what is often referred to as planning with a \emph{generative model}. We propose the AE-LSVI algorithm for best policy identification, a novel variant of the kernelized least-squares value iteration (LSVI) algorithm that combines optimism with pessimism for active exploration (AE). AE-LSVI provably identifies a near-optimal policy \emph{uniformly} over an entire state space and achieves polynomial sample complexity guarantees that are independent of the number of states. When specialized to the recently introduced offline contextual Bayesian optimization setting, our algorithm achieves improved sample complexity bounds. Experimentally, we demonstrate that AE-LSVI outperforms other RL algorithms in a variety of environments when robustness to the initial state is required.

****

Algorithmic Determination of the Combinatorial Structure of the Linear Regions of ReLU Neural Networks

Marissa Masden

We algorithmically determine the regions and facets of all dimensions of the canonical polyhedral complex, the universal object into which a ReLU network decomposes its input space. We show that the locations of the vertices of the canonical polyhedral complex along with their signs with respect to layer maps determine the full facet structure across all dimensions. We present an algorithm which calculates this full combinatorial structure, making use of our theorems that the dual complex to the canonical polyhedral complex is cubical and it possesses a multiplication compatible with its facet structure. The resulting algorithm is numerically stable, polynomial time in the number of intermediate neurons, and obtains accurate information across all dimensions. This permits us to obtain, for example, the true topology of the decision boundaries of networks with low-dime nsional inputs. We run empirics on such networks at initialization, finding that width alone does not increase observed topology, but width in the presence of depth does.

****

Are Neurons Actually Collapsed? On the Fine-Grained Structure in Neural Representations

Yongyi Yang,Jacob Steinhardt,Wei Hu

Recent work has observed an intriguing "Neural Collapse" phenomenon in well-trained neural networks, where the last-layer representations of training samples with the same label collapse into each other. This suggests that the last-layer representations are completely determined by the labels, and do not depend on the intrinsic structure of input distribution. We provide evidence that this is not a complete description, and that the apparent collapse hides important fine-grained structure in the representations. Specifically, even when representations apparently collapse, the small amount of remaining variation can still faithfully and accurately captures the intrinsic structure of input distribution. As an example, if we train on CIFAR-10 using only 5 coarse-grained labels (by combining two classes into one super-class) until convergence, we can reconstruct the original 10-class labels from the learned representations via unsupervised clustering. The reconstructed labels achieve 93% accuracy on the CIFAR-10 test set, nearly matching the normal CIFAR-10 accuracy for the same architecture. Our findings show concretely how the structure of input data can play a significant role in determining the fine-grained structure of neural representations, going beyond what Neural Collapse predicts.

****

FoSR: First-order spectral rewiring for addressing oversquashing in GNNs

Kedar Karhadkar,Pradeep Kr. Banerjee,Guido Montufar

Graph neural networks (GNNs) are able to leverage the structure of graph data by passing messages along the edges of the graph. While this allows GNNs to learn features depending on the graph structure, for certain graph topologies it leads to inefficient information propagation and a problem known as oversquashing. This has recently been linked with the curvature and spectral gap of the graph. On the other hand, adding edges to the message-passing graph can lead to increasingly similar node representations and a problem known as oversmoothing. We propose a computationally efficient algorithm that prevents oversquashing by systematically adding edges to the graph based on spectral expansion. We combine this with a relational architecture, which lets the GNN preserve the original graph structure and provably prevents oversmoothing. We find experimentally that our algorithm outperforms existing graph rewiring methods in several graph classification tasks.

****

Early Stopping for Deep Image Prior

Hengkang Wang,Taihui Li,Zhong Zhuang,Tiancong Chen,Hengyue Liang,Ju Sun

Deep image prior (DIP) and its variants have shown remarkable potential for solving inverse problems in computational imaging (CI), needing no separate training data. Practical DIP models are often substantially overparameterized. During the learning process, these models learn the desired visual content first and then pick up the potential modeling and observational noise, i.e., overfitting. Thus, the practicality of DIP hinges on early stopping (ES) that captures the transition period. In this regard, the majority of prior DIP works for CI tasks only demonstrate the potential of the models---reporting the peak performance against the groundtruth but providing no clue about how to operationally obtain near-peak performance without access to the groundtruth. In this paper, we set to break this practicality barrier of DIP, and propose an efficient ES strategy that consistently detects near-peak performance across several CI tasks and DIP variants. Simply based on the running variance of DIP intermediate reconstructions, our ES method not only outpaces the existing ones---which only work in very narrow regimes, but also remains effective when combined with methods that try to mitigate overfitting.

****

In-Context Policy Iteration

Ethan Brooks,Logan A Walls,Richard Lewis,Satinder Singh

This work presents In-Context Policy Iteration, an algorithm for performing Reinforcement Learning (RL), in-context, using foundation models. While the application of foundation models to RL has received considerable attention, most approaches rely on either (1) the curation of expert demonstrations (either through manual design or task-specific pretraining) or (2) adaptation to the task of interest using gradient methods (either fine-tuning or training of adapter layers). Both of these techniques have drawbacks. Collecting demonstrations is labor-intensive, and algorithms that rely on them do not outperform the experts from which the demonstrations were derived. All gradient techniques are inherently slow, sacrificing the “few-shot” quality that made in-context learning attractive to begin with. In this work, we present an algorithm, ICPI, that learns to perform RL tasks without expert demonstrations or gradients. Instead we present a policy-iteration method in which the prompt content is the entire locus of learning. ICPI iteratively updates the contents of the prompt from which it derives its policy through trial-and-error interaction with an RL environment. In order to eliminate the role of in-weights learning (on which approaches like Decision Transformer rely heavily), we demonstrate our algorithm using Codex Chen et al. (2021b), a language model with no prior knowledge of the domains on which we evaluate it.

****

Learning to Grow Pretrained Models for Efficient Transformer Training

Peihao Wang,Rameswar Panda,Lucas Torroba Hennigen,Philip Greengard,Leonid Karlinsky,Rogerio Feris,David Daniel Cox,Zhangyang Wang,Yoon Kim

Scaling transformers has led to significant breakthroughs in many domains, leading to a paradigm in which larger versions of existing models are trained and released on a periodic basis. New instances of such models are typically trained completely from scratch, despite the fact that they are often just scaled-up versions of their smaller counterparts. How can we use the implicit knowledge in the parameters of smaller, extant models to enable faster training of newer, larger models? This paper describes an approach for accelerating transformer training by learning to grow pretrained transformers, where we learn to linearly map the parameters of the smaller model to initialize the larger model. For tractable learning, we factorize the linear transformation as a composition of (linear) width- and depth-growth operators, and further employ a Kronecker factorization of these growth operators to encode architectural knowledge. Extensive experiments across both language and vision transformers demonstrate that our learned Linear Growth Operator (LiGO) can save up to 50% computational cost of training from scratch, while also consistently outperforming strong baselines that also reuse smaller pretrained models to initialize larger models.

****

Generative Modeling Helps Weak Supervision (and Vice Versa)

Benedikt Boecking,Nicholas Roberts,Willie Neiswanger,Stefano Ermon,Frederic Sala,Artur Dubrawski

Many promising applications of supervised machine learning face hurdles in the acquisition of labeled data in sufficient quantity and quality, creating an expensive bottleneck. To overcome such limitations, techniques that do not depend on ground truth labels have been studied, including weak supervision and generative modeling. While these techniques would seem to be usable in concert, improving one another, how to build an interface between them is not well-understood. In this work, we propose a model fusing programmatic weak supervision and generative adversarial networks and provide theoretical justification motivating this fusion. The proposed approach captures discrete latent variables in the data alongside the weak supervision derived label estimate. Alignment of the two allows for better modeling of sample-dependent accuracies of the weak supervision sources, improving the estimate of unobserved labels. It is the first approach to enable data augmentation through weakly supervised synthetic images and pseudolabels. Additionally, its learned latent variables can be inspected qualitatively. The model outperforms baseline weak supervision label models on a number of multiclass image classification datasets, improves the quality of generated images, and further improves end-model performance through data augmentation with synthetic samples.

****

What does a platypus look like? Generating customized prompts for zero-shot image classification

Sarah M Pratt,Rosanne Liu,Ali Farhadi

Open vocabulary models are a promising new paradigm for image classification. Unlike traditional classification models, open vocabulary models classify among any arbitrary set of categories specified with natural language during inference. This natural language, called "prompts", typically consists of a set of hand-written templates (e.g., "a photo of a {}”) which are completed with each of the category names. This work introduces a simple method to generate higher accuracy prompts, without relying on any explicit knowledge of the task domain and with far fewer hand-constructed sentences. To achieve this, we combine open vocabulary models with large language models (LLMs) to create Customized Prompts via Language models (CuPL, pronounced "couple"). In particular, we leverage the knowledge contained in LLMs in order to generate many descriptive sentences that are customized for each object category. We find that this straightforward and general approach improves accuracy on a range of zero-shot image classification benchmarks, including over one percentage point gain on ImageNet. Finally, this simple baseline requires no additional training and remains completely zero-shot.

****

Provable Memorization Capacity of Transformers

Junghwan Kim,Michelle Kim,Barzan Mozafari

Quantifying memorization capacity is essential for understanding the expressiveness and generalizability of deep learning model architectures. However, the memorization capacity of the Transformer architecture has yet to be explored. In this work, we present the first study of the memorization capacity of the Transformer architecture. We prove that Transformers are capable of memorizing N sequence-to-sequence mappings of length n with d-dimensional input tokens using $\tilde{O}(d + n + \sqrt{nN})$ parameters. Our theory supports memorization both with and without permutation equivariance, utilizing positional encodings in the latter case. Building on our theory, we also analyze the memorization capacity of Transformers in the sequence classification and language modeling tasks. To verify these theoretical findings, we conduct experiments analyzing the memorization capacity of Transformers in the natural language domain.

****

Knowledge-Driven New Drug Recommendation

Zhenbang Wu,Huaxiu Yao,Zhe Su,David M Liebovitz,Lucas M Glass,James Zou,Chelsea Finn,Jimeng Sun

Drug recommendation assists doctors in prescribing personalized medications to patients based on their health conditions. Existing drug recommendation solutions adopt the supervised multi-label classification setup and only work with existing drugs with sufficient prescription data from many patients. However, newly approved drugs do not have much historical prescription data and cannot leverage existing drug recommendation methods. To address this, we formulate the new drug recommendation as a few-shot learning problem. Yet, directly applying existing few-shot learning algorithms faces two challenges: (1) complex relations among diseases and drugs and (2) numerous false-negative patients who were eligible but did not yet use the new drugs. To tackle these challenges, we propose EDGE, which can quickly adapt to the recommendation for a new drug with limited prescription data from a few support patients. EDGE maintains a drug-dependent multi-phenotype few-shot learner to bridge the gap between existing and new drugs. Specifically, EDGE leverages the drug ontology to link new drugs to existing drugs with similar treatment effects and learns ontology-based drug representations. Such drug representations are used to customize the metric space of the phenotype-driven patient representations, which are composed of a set of phenotypes capturing complex patient health status. Lastly, EDGE eliminates the false-negative supervision signal using an external drug-disease knowledge base. We evaluate EDGE on two real-world datasets: the public EHR data (MIMIC-IV) and private industrial claims data. Results show that EDGE achieves 7.3% improvement on the ROC-AUC score over the best baseline.

****

Beyond Traditional Transfer Learning: Co-finetuning for Action Localisation

Anurag Arnab,Xuehan Xiong,Alexey A. Gritsenko,Rob Romijnders,Josip Djolonga,Mostafa Dehghani,Chen Sun,Mario Lucic,Cordelia Schmid

Transfer learning is the predominant paradigm for training deep networks on small target datasets. Models are typically pretrained on large “upstream” datasets for classification, as such labels are easy to collect, and then finetuned on downstream” tasks such as action localisation, which are smaller due to their finer-grained annotations.

In this paper, we question this approach, and propose co-finetuning -- simultaneously training a single model on multiple “upstream” and “downstream” tasks. We demonstrate that co-finetuning outperforms traditional transfer learning when using the same total amount of data, and also show how we can easily extend our approach to multiple “upstream” datasets to further improve performance. In particular, co-finetuning significantly improves the performance on rare classes in our downstream task, as it has a regularising effect, and enables the network to learn feature representations that transfer between different datasets. Finally, we observe how co-finetuning with public, video classification datasets, we are able to achieve state-of-the-art results for spatio-temporal action localisation on the challenging AVA and AVA-Kinetics datasets, outperforming recent works which develop intricate models.

****

Output Distribution over the Entire Input Space: A Novel Perspective to Understand Neural Networks

Weitang Liu,Yi-Zhuang You,Jingbo Shang

Understanding the input-output mapping relationship in the \emph{entire input space} contributes a novel perspective to a comprehensive understanding of deep neural networks. In this paper, we focus on binary neural classifiers and propose to first uncover the histogram about the number of inputs that are mapped to certain output values and then scrutinize the representative inputs from a certain output range of interest, such as the positive-logit region that corresponds to one of the classes. A straightforward solution is uniform sampling (or exhaustive enumeration) in the entire input space but when the inputs are high dimensional, it can take almost forever to converge. We connect the output histogram to the \emph{density of states} in physics by making an analogy between the energy of a system and the neural network output. Inspired by the Wang-Landau algorithm designed for sampling the density of states, we propose an efficient sampler that is driven to explore the under-explored output values through a gradient-based proposal. Compared with the random proposal in Wang-Landau algorithm, our gradient-based proposal converges faster as it can propose the inputs corresponding to the under-explored output values. Extensive experiments have verified the accuracy of the histogram generated by our sampler and also demonstrated interesting findings. For example, the models map many human unrecognizable images to very negative logit values. These properties of a neural model are revealed for the first time through our sampled statistics. We believe that our approach opens a new gate for neural model evaluation and shall be further explored in future works.

****

Learning Control Policies for Region Stabilization in Stochastic Systems

Matin Ansaripour,Mathias Lechner,■or■e Žikeli■,Krishnendu Chatterjee,Thomas A Henzinger

We consider the problem of learning control policies in stochastic systems which guarantee that the system stabilizes within some specified stabilization region with probability 1. Our approach is based on the novel notion of stabilizing ranking supermartingales (sRSMs) that we introduce in this work. Our sRSMs overcom

e the limitation of methods proposed in previous works whose applicability is restricted to systems in which the stabilizing region cannot be left once entered under any control policy. We present a learning procedure that learns a control policy together with an sRSM that formally certifies probability 1 stability, both learned as neural networks. Our experimental evaluation shows that our learning procedure can successfully learn provably stabilizing policies in practice.
****

InCoder: A Generative Model for Code Infilling and Synthesis
Daniel Fried,Armen Aghajanyan,Jessy Lin,Sida Wang,Eric Wallace,Freda Shi,Ruiqi Zhong,Scott Yih,Luke Zettlemoyer,Mike Lewis
Code is seldom written in a single left-to-right pass and is instead repeatedly edited and refined. We introduce InCoder, a unified generative model that can perform program synthesis (via left-to-right generation) as well as editing (via masking and infilling). InCoder is trained to generate code files from a large corpus of permissively licensed code, where regions of code have been randomly masked and moved to the end of each file, allowing code infilling with bidirectional context. Our model is the first large generative code model that is able to infill arbitrary regions of code, which we evaluate in a zero-shot setting on challenging tasks such as type inference, comment generation, and variable re-naming. We find that the ability to condition on bidirectional context substantially improves performance on these tasks, while still performing comparably on standard program synthesis benchmarks in comparison to left-to-right only models pretrained at similar scale. Our models and code will be publicly released.
****

Bridge the Inference Gaps of Neural Processes via Expectation Maximization
Qi Wang,Marco Federici,Herke van Hoof
The neural process (NP) is a family of computationally efficient models for learning distributions over functions. However, it suffers from under-fitting and shows suboptimal performance in practice. Researchers have primarily focused on incorporating diverse structural inductive biases, e.g. attention or convolution, in modeling. The topic of inference suboptimality and an analysis of the NP from the optimization objective perspective has hardly been studied in earlier work. To fix this issue, we propose a surrogate objective of the target log-likelihood of the meta dataset within the expectation maximization framework. The resulting model, referred to as the Self-normalized Importance weighted Neural Process (SI-NP), can learn a more accurate functional prior and has an improvement guarantee concerning the target log-likelihood. Experimental results show the competitive performance of SI-NP over other NPs objectives and illustrate that structural inductive biases, such as attention modules, can also augment our method to achieve SOTA performance.
****

Generated Graph Detection
Yihan Ma,Zhikun Zhang,Ning Yu,Xinlei He,Michael Backes,Yun Shen,Yang Zhang
Graph generative models become increasingly effective for data distribution approximation and data augmentation. Although still in sandboxes, they have aroused public concerns about their malicious misuses or misinformation broadcasts, just as what Deepfake visual and auditory media has been delivering to society. It is never too early to regulate the prevalence of generated graphs. As a preventive response, we pioneer to formulate the generated graph detection problem to distinguish generated graphs from real ones. We propose the first framework to systematically investigate a set of sophisticated models and their performance in four classification scenarios. Each scenario switches between seen and unseen datasets/generators during testing to get closer to real world settings and progressively challenge the classifiers. Extensive experiments evidence that all the models are qualified for generated graph detection, with specific models having advantages in specific scenarios. Resulting from the validated generality and oblivion of the classifiers to unseen datasets/generators, we draw a safe conclusion that our solution can sustain for a decent while to curb generated graph misuses.
****

Neural Embeddings for Text
Oleg Vasilyev,John Bohannon

We propose a new kind of embedding for natural language text that deeply represents semantic meaning. Standard text embeddings use the vector output of a pretrained language model. In our method, we let a language model learn from the text and then literally pick its brain, taking the actual weights of the model's neurons to generate a vector. We call this representation of the text a neural embedding. With analysis of its behavior on several datasets, we confirm the ability of this representation to reflect semantics of the text. We also compare neural embeddings with GPT sentence (SGPT) embeddings. We observe that neural embeddings achieve comparable performance with a far smaller model, and that the embeddings respond to semantics differently.

****

Find Your Friends: Personalized Federated Learning with the Right Collaborators
Yi Sui,Junfeng Wen,Yenson Lau,Brendan Leigh Ross,Jesse C Cresswell

In the traditional federated learning setting, a central server coordinates a network of clients to train one global model. However, the global model may serve many clients poorly due to data heterogeneity. Moreover, there may not exist a trusted central party that can coordinate the clients to ensure that each of them can benefit from others. To address these concerns, we present a novel decentralized framework, FedeRiCo, where each client can learn as much or as little from other clients as is optimal for its local data distribution. Based on expectation-maximization, FedeRiCo estimates the utilities of other participants’ models on each client’s data so that everyone can select the right collaborators for learning. As a result, our algorithm outperforms other federated, personalized, and/or decentralized approaches on several benchmark datasets, being the only approach that consistently performs better than training with local data only.

****

Masked Vision and Language Modeling for Multi-modal Representation Learning
Gukyeong Kwon,Zhaowei Cai,Avinash Ravichandran,Erhan Bas,Rahul Bhotika,Stefano Soatto

In this paper, we study how to use masked signal modeling in vision and language (V+L) representation learning. Instead of developing masked language modeling (MLM) and masked image modeling (MIM) independently, we propose to build joint masked vision and language modeling, where the masked signal of one modality is reconstructed with the help from another modality. This is motivated by the nature of image-text paired data that both of the image and the text convey almost the same information but in different formats. The masked signal reconstruction of one modality conditioned on another modality can also implicitly learn cross-modal alignment between language tokens and image patches. Our experiments on various V+L tasks show that the proposed method, along with common V+L alignment losses, not only achieves state-of-the-art performance by using a large amount of data but also outperforms the other competitors by a significant margin in the regimes of limited training data.

****

Quantum Fourier Networks for solving Parametric PDEs
Nishant Jain,Jonas Landman,Natansh Mathur,Iordanis Kerenidis

Many real-world problems like modelling environment dynamics, physical processes, time series etc., involve solving Partial Differential Equations (PDEs) parameterized by problem-specific conditions. Recently, a deep learning architecture called Fourier Neural Operator (FNO) proved to be capable of learning solutions of given PDE families, for any initial conditions as input. Given the advancements in quantum hardware and the recent results in quantum machine learning methods, we propose three quantum circuits, inspired by the FNO, to learn this functional mapping for PDEs. The proposed algorithms are distinguished based on the trade-off between depth and their similarity to the classical FNO. At their core, we make use of unary encoding paradigm and orthogonal quantum layers, and introduce a new quantum Fourier transform in the unary basis. With respect to the number of samples, our quantum algorithm is proven to be substantially faster than the classical counterpart. We benchmark our proposed algorithms on three PDE famili es, namely Burger’s equation, Darcy’s flow equation and the Navier-Stokes equation, and the results show that our quantum methods are comparable in performance to the classical FNO. We also show an analysis of the image classification tasks where our proposed algorithms are able to match the accuracy of the CNNs, thereby showing their applicability to other domains.

****

Agent-based Graph Neural Networks
Karolis Martinkus,Pál András Papp,Benedikt Schesch,Roger Wattenhofer

We present a novel graph neural network we call AgentNet, which is designed specifically for graph-level tasks. AgentNet is inspired by sublinear algorithms, featuring a computational complexity that is independent of the graph size. The architecture of AgentNet differs fundamentally from the architectures of traditional graph neural networks. In AgentNet, some trained \textit{neural agents} intelligently walk the graph, and then collectively decide on the output. We provide an extensive theoretical analysis of AgentNet: We show that the agents can learn to systematically explore their neighborhood and that AgentNet can distinguish some structures that are even indistinguishable by 2-WL. Moreover, AgentNet is able to separate any two graphs which are sufficiently different in terms of subgraphs. We confirm these theoretical results with synthetic experiments on hard-to-distinguish graphs and real-world graph classification tasks. In both cases, we compare favorably not only to standard GNNs but also to computationally more expensive GNN extensions.

****

Generating Adversarial Examples with Task Oriented Multi-Objective Optimization
Anh Tuan Bui,Trung Le,He Zhao,Quan Hung Tran,Paul Montague,Dinh Phung

Deep learning models, even the-state-of-the-art ones, are highly vulnerable to adversarial examples. Adversarial training is one of the most efficient methods to improve the model's robustness. The key factor for the success of adversarial training is the capability to generate qualified and divergent adversarial examples which satisfy some objectives/goals (e.g., finding adversarial examples that maximize the model losses for simultaneously attacking multiple models). Therefore, multi-objective optimization (MOO) is a natural tool for adversarial example generation, where we search adversarial examples simultaneously maximizing some objectives/goals. However, we observe that a naive application of MOO tends to maximize all objectives/goals equally, without caring if an objective/goal has been achieved yet. This leads to useless effort to further improve the goal-achieved tasks, while putting less focus on the goal-unachieved tasks. In this paper, we propose \emph{Task Oriented MOO} to address this issue, in the context where we can explicitly define the goal achievement for a task. Our principle is to only maintain the goal-achieved tasks, while letting the optimizer spend more effort on improving the goal-unachieved tasks. We conduct comprehensive experiments for our Task Oriented MOO on various adversarial example generation schemes. The experimental results firmly demonstrate the merit of our proposed approach.

****

On the Performance of Temporal Difference Learning With Neural Networks
HAOXING TIAN,Ioannis Paschalidis,Alex Olshevsky

Neural Temporal Difference (TD) Learning is an approximate temporal difference method for policy evaluation that uses a neural network for function approximation. Analysis of Neural TD Learning has proven to be challenging. In this paper we provide a convergence analysis of Neural TD Learning with a projection onto $B(\theta_0, \omega)$, a ball of fixed radius ω around the initial point θ_0. We show an approximation bound of $O(\epsilon + 1/\sqrt{m})$ where ϵ is the approximation quality of the best neural network in $B(\theta_0, \omega)$ and m is the width of all hidden layers in the network.

****

Certified Defences Against Adversarial Patch Attacks on Semantic Segmentation
Maksym Yatsura,Kaspar Sakmann,N. Grace Hua,Matthias Hein,Jan Hendrik Metzen

Adversarial patch attacks are an emerging security threat for real world deep learning applications. We present Demasked Smoothing, the first approach (up to our knowledge) to certify the robustness of semantic segmentation models against t his threat model. Previous work on certifiably defending against patch attacks has mostly focused on image classification task and often required changes in the model architecture and additional training which is undesirable and computationally expensive. In Demasked Smoothing, any segmentation model can be applied without particular training, fine-tuning, or restriction of the architecture. Using different masking strategies, Demasked Smoothing can be applied both for certified detection and certified recovery. In extensive experiments we show that Demasked Smoothing can on average certify 63% of the pixel predictions for a 1% patch in the detection task and 46% against a 0.5% patch for the recovery task on the ADE20K dataset.

****

Markup-to-Image Diffusion Models with Scheduled Sampling

Yuntian Deng,Noriyuki Kojima,Alexander M Rush

Building on recent advances in image generation, we present a fully data-driven approach to rendering markup into images. The approach is based on diffusion models, which parameterize the distribution of data using a sequence of denoising operations on top of a Gaussian noise distribution. We view the diffusion denoising process a sequential decision making process, and show that it exhibits compounding errors similar to exposure bias issues in imitation learning problems. To mitigate these issues, we adapt the scheduled sampling algorithm to diffusion training. We conduct experiments on four markup datasets: formulas (LaTeX), table layouts (HTML), sheet music (LilyPond), and molecular images (SMILES). These experiments each verify the effectiveness of diffusion and the use of scheduled sampling to fix generation issues. These results also show that the markup-to-image task presents a useful controlled compositional setting for diagnosing and analyzing generative image models.

****

Efficient Reward Poisoning Attacks on Online Deep Reinforcement Learning

Yinglun Xu,Qi Zeng,Gagandeep Singh

We study reward poisoning attacks on online deep reinforcement learning (DRL), where the attacker is oblivious to the learning algorithm used by the agent and does not necessarily have full knowledge of the environment. We demonstrate the intrinsic vulnerability of state-of-the-art DRL algorithms by designing a general, black-box reward poisoning framework called adversarial MDP attacks. We instantiate our framework to construct several new attacks which only corrupt the rewards for a small fraction of the total training timesteps and make the agent learn a low-performing policy. Our key insight is that state-of-the-art DRL algorithms strategically explore the environment to find a high-performing policy. Our attacks leverage this insight to construct a corrupted environment where (a) the agent learns a high-performing policy that has low performance in the original environment and (b) the corrupted environment is similar to the original one so that the attacker's budget is reduced. We provide a theoretical analysis of the efficiency of our attack and perform an extensive evaluation. Our results show that our attacks efficiently poison agents learning with a variety of state-of-the-art DRL algorithms, such as DQN, PPO, SAC, etc., under several popular classical control and MuJoCo environments.

****

How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules

Yutong Xie,Ziqiao Xu,Jiaqi Ma,Qiaozhu Mei

Forming a molecular candidate set that contains a wide range of potentially effective compounds is crucial to the success of drug discovery. While most databases and machine-learning-based generation models aim to optimize particular chemical properties, there is limited literature on how to properly measure the coverage of the chemical space by those candidates included or generated. This problem is challenging due to the lack of formal criteria to select good measures of the chemical space. In this paper, we propose a novel evaluation framework for measures of the chemical space based on two analyses: an axiomatic analysis with three intuitive axioms that a good measure should obey, and an empirical analysis on the correlation between a measure and a proxy gold standard. Using this frame work, we are able to identify #Circles, a new measure of chemical space coverage, which is superior to existing measures both analytically and empirically. We further evaluate how well the existing databases and generation models cover the chemical space in terms of #Circles. The results suggest that many generation models fail to explore a larger space over existing databases, which leads to new opportunities for improving generation models by encouraging exploration.

****

D-CIPHER: Discovery of Closed-form Partial Differential Equations
Krzysztof Kacprzyk,Zhaozhi Qian,Mihaela van der Schaar
Closed-form differential equations, including partial differential equations and higher-order ordinary differential equations, are one of the most important tools used by scientists to model and better understand natural phenomena. Discovering these equations directly from data is challenging because it requires modeling relationships between various derivatives that are not observed in the data (equation-data mismatch) and it involves searching across a huge space of possible equations. Current approaches make strong assumptions about the form of the equation and thus fail to discover many well-known systems. Moreover, many of them resolve the equation-data mismatch by estimating the derivatives, which makes them inadequate for noisy and infrequently sampled systems. To this end, we propose D-CIPHER, which is robust to measurement artifacts and can uncover a new and very general class of differential equations. We further design a novel optimization procedure, CoLLie, to help D-CIPHER search through this class efficiently. Finally, we demonstrate empirically that it can discover many well-known equations that are beyond the capabilities of current methods.

****

Towards Identification of Microaggressions in real-life and Scripted conversations, using Context-Aware Machine Learning Techniques.
Mikel K. Ngueajio,Ivan Hernandez,Kelsi Cornett,Gloria Washington,darryl parsons
The advent and rapid proliferation of social media have brought with it an exponential growth in hate speech and overt offensive language, with one of the most subtle yet pervasive subcategories of hate speech being Microaggressions (MA). MAs are unintentional, hostile, derogatory, or negative prejudicial slights and insults toward any group, particularly culturally marginalized communities and growing bodies of research are linking long-term MA exposure to serious health problems. The scarcity of studies leveraging AI techniques to identify MAs in text and in spoken conversations, coupled with the lack of investigative analysis on the impact of context on the performance of algorithms used for this task, makes this a relevant topic for the AI community. In this paper, we explore the degree of effectiveness of MAs detection often found in spoken human communications across various contexts (e.g., workplace, social media, conversations) using Machine Learning models. We further examine the extent that art may imitate life, by comparing the ability of these models trained on real-life conversations to infer MAs, occurring in scripted Television shows. We apply a Support Vector Machine (SVM) classifier using N-grams and contextual modeling representation, using the Robustly Optimized Bidirectional Encoder Representation for Transformer (RoBERTa) model, whose performance is evaluated based on its pretraining size and ability to accurately detect hate speeches, with comparative results from BERT based-uncased and the HateBERT model respectively.
Overall, the results show that contextual transformer models outperform simpler context-free approaches to classifying MAs collected from surveys and online blogs. We also found that these models trained on real-life conversations could infer MAs in scripted TV settings, though at reduced levels, and equal rates, suggesting there may be a disconnect between contexts of MA found in art and those from real life.

****

UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks
Jiasen Lu,Christopher Clark,Rowan Zellers,Roozbeh Mottaghi,Aniruddha Kembhavi
We propose Unified-IO, a model that performs a large variety of AI tasks spanning classical computer vision tasks, including pose estimation, object detection,

depth estimation and image generation, vision-and-language tasks such as region captioning and referring expression, to natural language processing tasks such as question answering and paraphrasing. Developing a single unified model for such a large variety of tasks poses unique challenges due to the heterogeneous inputs and outputs pertaining to each task, including RGB images, per-pixel maps, binary masks, bounding boxes, and language. We achieve this unification by homogenizing every supported input and output into a sequence of discrete vocabulary tokens. This common representation across all tasks allows us to train a single transformer-based architecture, jointly on over 90 diverse datasets in the vision and language fields. Unified-IO is the first model capable of performing all 7 tasks on the GRIT benchmark and produces strong results across 16 diverse benchmarks like NYUv2-Depth, ImageNet, VQA2.0, OK-VQA, Swig, VizWizGround, BoolQ, and SciTail, with no task-specific fine-tuning. Code and pre-trained models will be made publicly available.

****

Benchmarking Offline Reinforcement Learning on Real-Robot Hardware

Nico Gürtler,Sebastian Blaes,Pavel Kolev,Felix Widmaier,Manuel Wuthrich,Stefan Bauer,Bernhard Schölkopf,Georg Martius

Learning policies from previously recorded data is a promising direction for real-world robotics tasks, as online learning is often infeasible. Dexterous manipulation in particular remains an open problem in its general form. The combination of offline reinforcement learning with large diverse datasets, however, has the potential to lead to a breakthrough in this challenging domain analogously to the rapid progress made in supervised learning in recent years. To coordinate the efforts of the research community toward tackling this problem, we propose a benchmark including: i) a large collection of data for offline learning from a dexterous manipulation platform on two tasks, obtained with capable RL agents trained in simulation; ii) the option to execute learned policies on a real-world robotic system and a simulation for efficient debugging. We evaluate prominent open-sourced offline reinforcement learning algorithms on the datasets and provide a reproducible experimental setup for offline reinforcement learning on real systems.

****

CUDA: Curriculum of Data Augmentation for Long-tailed Recognition

Sumyeong Ahn,Jongwoo Ko,Se-Young Yun

Class imbalance problems frequently occur in real-world tasks, and conventional deep learning algorithms are well known for performance degradation on imbalanced training datasets. To mitigate this problem, many approaches have aimed to balance among given classes by re-weighting or re-sampling training samples. These re-balancing methods increase the impact of minority classes and reduce the influence of majority classes on the output of models. However, the extracted representations may be of poor quality owing to the limited number of minority samples. To handle this restriction, several methods have been developed that increase the representations of minority samples by leveraging the features of the majority samples. Despite extensive recent studies, no deep analysis has been conducted on determination of classes to be augmented and strength of augmentation has been conducted. In this study, we first investigate the correlation between the degree of augmentation and class-wise performance, and find that the proper degree of augmentation must be allocated for each class to mitigate class imbalance problems. Motivated by this finding, we propose a simple and efficient novel curriculum, which is designed to find the appropriate per-class strength of data augmentation, called CUDA: CUrriculum of Data Augmentation for long-tailed recognition. CUDA can simply be integrated into existing long-tailed recognition methods. We present the results of experiments showing that CUDA effectively achieves better generalization performance compared to the state-of-the-art method on various imbalanced datasets such as CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018.

****

Understanding new tasks through the lens of training data via exponential tiltin g

Subha Maity,Mikhail Yurochkin,Moulinath Banerjee,Yuekai Sun

Deploying machine learning models on new tasks is a major challenge due to differences in distributions of the train (source) data and the new (target) data. However, the training data likely captures some of the properties of the new task. We consider the problem of reweighing the training samples to gain insights into the distribution of the target task. Specifically, we formulate a distribution shift model based on the exponential tilt assumption and learn train data importance weights minimizing the KL divergence between labeled train and unlabeled target datasets. The learned train data weights can then be used for downstream tasks such as target performance evaluation, fine-tuning, and model selection. We demonstrate the efficacy of our method on Waterbirds and Breeds benchmarks.

****

Neighborhood Gradient Clustering: An Efficient Decentralized Learning Method for Non-IID Data Distributions

Sai Aparna Aketi,Sangamesh Kodge,Kaushik Roy

Decentralized learning algorithms enable the training of deep learning models over large distributed datasets generated at different devices and locations, without the need for a central server. In practical scenarios, the distributed datasets can have significantly different data distributions across the agents. The current state-of-the-art decentralized algorithms mostly assume the data distributions to be Independent and Identically Distributed (IID). This paper focuses on improving decentralized learning over non-IID data distributions with minimal compute and memory overheads. We propose Neighborhood Gradient Clustering (NGC), a novel decentralized learning algorithm that modifies the local gradients of each agent using self- and cross-gradient information. Cross-gradients for a pair of neighboring agents are the derivatives of the model parameters of an agent with respect to the dataset of the other agent. In particular, the proposed method replaces the local gradients of the model with the weighted mean of the self-gradients, model-variant cross-gradients (derivatives of the received neighbors’ model parameters with respect to the local dataset - computed locally), and data-variant cross-gradients (derivatives of the local model with respect to its neighbors’ datasets - received through communication). The data-variant cross-gradients are aggregated through an additional communication round without breaking the privacy constraints of the decentralized setting. Further, we present CompNGC, a compressed version of NGC that reduces the communication overhead by $32 \times$ by compressing the cross-gradients. We demonstrate the empirical convergence and efficiency of the proposed technique over non-IID data distributions sampled from the CIFAR-10 dataset on various model architectures and graph topologies. Our experiments demonstrate that NGC and CompNGC outperform the existing state-of-the-art (SoTA) decentralized learning algorithm over non-IID data by $1-5\%$ with significantly less compute and memory requirements. Further, we also show that the proposed NGC method outperforms the baseline by $5-40\%$ with no additional communication.

****

Equilibrium-finding via exploitability descent with learned best-response functions

Carlos Martin,Tuomas Sandholm

There has been great progress on equilibrium-finding research over the last 20 years. Most of that work has focused on games with finite, discrete action spaces. However, many games involving space, time, money, etc. have continuous action spaces. We study the problem of computing approximate Nash equilibria of games with continuous strategy sets. The main measure of closeness to Nash equilibrium is exploitability, which measures how much players can benefit from unilaterally changing their strategy. We propose a new method that minimizes an approximation of exploitability with respect to the strategy profile. This approximation is computed using learned best-response functions, which take the current strategy profile as input and return learned best responses. The strategy profile and best-response functions are trained simultaneously, with the former trying to minimize exploitability while the latter try to maximize it. We evaluate our method on various continuous games, showing that it outperforms prior methods.

****

A Unified Framework for Comparing Learning Algorithms

Harshay Shah,Sung Min Park,Andrew Ilyas,Aleksander Madry

We propose a framework for {\em (learning) algorithm comparisons}, wherein the goal is to find similarities and differences between models trained with two different learning algorithms. We begin by formalizing the goal of algorithm comparison as finding {\em distinguishing feature transformations}, input transformations that change the predictions of models trained with one learning algorithm but not the other. We then present a two-stage method for algorithm comparisons based on comparing how models use the training data, leveraging the recently proposed datamodel representations [Ilyas et al., 2022]. We demonstrate our framework through three case studies that compare models trained with/without standard data augmentation, with/without pre-training, and with different optimizer hyperparameters.

****

Neural Network Approximations of PDEs Beyond Linearity: Representational Perspective

Tanya Marwah,Zachary Chase Lipton,Jianfeng Lu,Andrej Risteski

A burgeoning line of research has developed deep neural networks capable of approximating the solutions to high dimensional PDEs, opening related lines of theoretical inquiry focused on explaining how it is that these models appear to evade the curse of dimensionality. However, most theoretical analyses thus far have been limited to simple linear PDEs. In this work, we take a step towards studying the representational power of neural networks for approximating solutions to nonlinear PDEs. We focus on a class of PDEs known as nonlinear elliptic variational PDEs, whose solutions minimize an Euler-Lagrange energy functional $\mathcal{E}(u) = \int_\Omega L(\nabla u) dx$. We show that if composing a function with Barron norm b with L produces a function of Barron norm at most $B_L b^p$, the solution to the PDE can be ϵ-approximated in the L^2 sense by a function with Barron norm $O\left(\left(dB_L\right)^{p^{\log(1/\epsilon)}}\right)$. By a classical result due to \cite{barron1993universal}, this correspondingly bounds the size of a 2-layer neural network needed to approximate the solution. Treating p, ϵ, B_L as constants, this quantity is polynomial in dimension, thus showing neural networks can evade the curse of dimensionality. Our proof technique involves neurally simulating (preconditioned) gradient in an appropriate Hilbert space, which converges exponentially fast to the solution of the PDE, and such that we can bound the increase of the Barron norm at each iterate. Our results subsume and substantially generalize analogous prior results for linear elliptic PDEs.

****

Calibrating Sequence likelihood Improves Conditional Language Generation

Yao Zhao,Mikhail Khalman,Rishabh Joshi,Shashi Narayan,Mohammad Saleh,Peter J Liu

Conditional language models are predominantly trained with maximum likelihood estimation (MLE), giving probability mass to sparsely observed target sequences. While MLE trained models assign high probability to plausible sequences given the context, the model probabilities often do not accurately rank-order generated sequences by quality. This has been empirically observed in beam search decoding as output quality degrading with large beam sizes, and decoding strategies benefiting from heuristics such as length normalization and repetition-blocking. In this work, we introduce sequence likelihood calibration (SLiC) where the likelihood of model generated sequences are calibrated to better align with reference sequences in the model’s latent space. With SLiC, decoding heuristics become unnecessary and decoding candidates’ quality significantly improves regardless of the decoding method. Furthermore, SLiC shows no sign of diminishing returns with model scale, and presents alternative ways to improve quality with limited training and inference budgets. With SLiC, we exceed or match SOTA results on a wide range of generation tasks spanning abstractive summarization, question generation, abstractive question answering and data-to-text generation, even with modest-sized models.

****

Masked inverse folding with sequence transfer for protein representation learning

Kevin K Yang,Hugh Yeh,Niccolò Zanichelli

Self-supervised pretraining on protein sequences has led to state-of-the art performance on protein function and fitness prediction.
However, sequence-only methods ignore the rich information contained in experimental and predicted protein structures.
Meanwhile, inverse folding methods reconstruct a protein's amino-acid sequence given its structure, but do not take advantage of sequences that do not have known structures.
In this study, we train a masked inverse folding protein language model parameterized as a structured graph neural network.
We then show that using the outputs from a pretrained sequence-only protein masked language model as input to the inverse folding model further improves pretraining perplexity.
We evaluate both of these models on downstream protein engineering tasks and analyze the effect of using information from experimental or predicted structures on performance.

****

Convolutions are competitive with transformers for protein sequence pretraining

Kevin K Yang,Alex Xijie Lu,Nicolo Fusi

Pretrained protein sequence language models largely rely on the transformer architecture. However, transformer run-time and memory requirements scale quadratically with sequence length. We investigate the potential of a convolution-based architecture for protein sequence masked language model pretraining and subsequent finetuning. CNNs are competitive on the pretraining task with transformers across several orders of magnitude in parameter size while scaling linearly with sequence length. More importantly, CNNs are competitive with and occasionally superior to transformers across an extensive set of downstream evaluations, including structure prediction, zero-shot mutation effect prediction, and out-of-domain generalization. We also demonstrate strong performance on sequences longer than the positional embeddings allowed in the current state-of-the-art transformer protein masked language models. Finally, we close with a call to disentangle the effects of pretraining task and model architecture when studying pretrained protein sequence models.

****

Learning differentiable solvers for systems with hard constraints

Geoffrey Négiar,Michael W. Mahoney,Aditi Krishnapriyan

We introduce a practical method to enforce partial differential equation (PDE) constraints for functions defined by neural networks (NNs), up to a desired tolerance. By combining methods in differentiable optimization and applications of the implicit function theorem to NN models, we develop a differentiable PDE-constrained layer that can be incorporated into a NN. Inspired by dictionary learning, our model learns a family of functions, each of which defines a mapping from PDE parameters to PDE solutions. At inference time, the model finds an optimal linear combination of the functions in the learned family by solving a PDE-constrained optimization problem. Our method provides continuous solutions over the domain of interest that accurately satisfy desired physical constraints. Our results show that incorporating hard constraints directly into the NN architecture achieves much lower test error when compared to training on an unconstrained objective.

****

FedDAR: Federated Domain-Aware Representation Learning

Aoxiao Zhong,Hao He,Zhaolin Ren,Na Li,Quanzheng Li

Cross-silo Federated learning (FL) has become a promising tool in machine learning applications for healthcare. It allows hospitals/institutions to train models with sufficient data while the data is kept private. To make sure the FL model is robust when facing heterogeneous data among FL clients, most efforts focus on personalizing models for clients. However, the latent relationships between cli ents' data are ignored. In this work, we focus on a special non-iid FL problem, called Domain-mixed FL, where each client's data distribution is assumed to be a mixture of several predefined domains. Recognizing the diversity of domains and the similarity within domains, we propose a novel method, FedDAR, which learns a domain shared representation and domain-wise personalized prediction heads in a decoupled manner. For simplified linear regression settings, we have theoretically proved that FedDAR enjoys a linear convergence rate. For general settings, we have performed intensive empirical studies on both synthetic and real-world medical datasets which demonstrate its superiority over prior FL methods. Our code is available at https://github.com/zlz0414/FedDAR.

****

Learning to Estimate Shapley Values with Vision Transformers

Ian Connick Covert,Chanwoo Kim,Su-In Lee

Transformers have become a default architecture in computer vision, but understanding what drives their predictions remains a challenging problem. Current explanation approaches rely on attention values or input gradients, but these provide a limited view of a model’s dependencies. Shapley values offer a theoretically sound alternative, but their computational cost makes them impractical for large, high-dimensional models. In this work, we aim to make Shapley values practical for vision transformers (ViTs). To do so, we first leverage an attention masking approach to evaluate ViTs with partial information, and we then develop a procedure to generate Shapley value explanations via a separate, learned explainer model. Our experiments compare Shapley values to many baseline methods (e.g., attention rollout, GradCAM, LRP), and we find that our approach provides more accurate explanations than existing methods for ViTs.

****

No Double Descent in PCA: Training and Pre-Training in High Dimensions

Daniel Gedon,Antonio H. Ribeiro,Thomas B. Schön

With the recent body of work on overparameterized models the gap between theory and practice in contemporary machine learning is shrinking. While many of the present state-of-the-art models have an encoder-decoder architecture, there is little theoretical work for this model structure. To improve our understanding in this direction, we consider linear encoder-decoder models, specifically PCA with linear regression on data from a low-dimensional manifold. We present an analysis for fundamental guarantees of the risk and asymptotic results for isotropic data when the model is trained in a supervised manner. The results are also verified in simulations. Furthermore, we extend our analysis to the popular setting where parts of the model are pre-trained in an unsupervised manner by pre-training the PCA encoder with subsequent supervised training of the linear regression. We show that the overall risk depends on the estimates of the eigenvectors in the encoder and present a sample complexity requirement through a concentration bound. The results highlight that using more pre-training data decreases the overall risk only if it improves the eigenvector estimates. Therefore, we stress that the eigenvalue distribution determines whether more pre-training data is useful or not.

****

Predicting Drug Repurposing Candidates and Their Mechanisms from A Biomedical Knowledge Graph

Chunyu Ma,Zhihan Zhou,Han Liu,David Koslicki

Computational drug repurposing is a cost- and time-efficient method to identify new indications of approved or experimental drugs/compounds. It is especially critical for emerging and/or orphan diseases due to its cheaper investment and shorter research cycle compared with traditional wet-lab drug discovery approaches. However, the underlying mechanisms of action between repurposed drugs and their target diseases remain largely unknown, which is still an unsolved issue in existing repurposing methods. As such, computational drug repurposing has not been widely adopted in clinical settings. In this work, based on a massive biomedical knowledge graph, we propose a computational drug repurposing framework that not only predicts the treatment probabilities between drugs and diseases but also predicts the path-based, testable mechanisms of action (MOAs) as their biomedical

explanations. Specifically, we utilize the GraphSAGE model in an unsupervised manner to integrate each entity’s neighborhood information and employ a Random Forest model to predict the treatment probabilities between pairs of drugs and diseases. Moreover, we train an adversarial actor-critic reinforcement learning model to predict the potential MOA for explaining drug purposing. To encourage the model to find biologically reasonable paths, we utilize the curated molecular interactions of drugs and a PubMed-publication-based concept distance to extract potential drug MOA paths from the knowledge graph as ”demonstration paths” to guide the model during the process of path-finding. Comprehensive experiments and case studies show that the proposed framework outperforms state-of-the-art baselines in both predictive performance of drug repurposing and explanatory performance of recapitulating human-curated DrugMechDB-based paths.
****

Interval Bound Interpolation for Few-shot Learning with Few Tasks

Shounak Datta,Sankha Subhra Mullick,Anish Chakrabarty,Swagatam Das

Few-shot learning aims to transfer the knowledge acquired from training on a diverse set of tasks to unseen tasks from the same task distribution, with a limited amount of labeled data. The underlying requirement for effective few-shot generalization is to learn a good representation of the task manifold. This becomes more difficult when only a limited number of tasks are available for training. In such a few-task few-shot setting, it is beneficial to explicitly preserve the local neighborhoods from the task manifold and exploit this to generate artificial tasks for training. To this end, we introduce the notion of interval bounds from the provably robust training literature to few-shot learning. The interval bounds are used to characterize neighborhoods around the training tasks. These neighborhoods can then be preserved by minimizing the distance between a task and its respective bounds. We then use a novel strategy to artificially form new tasks for training by interpolating between the available tasks and their respective interval bounds. We apply our framework to both model-agnostic meta-learning as well as prototype-based metric-learning paradigms. The efficacy of our proposed approach is evident from the improved performance on several datasets from diverse domains in comparison to recent methods.
****

A framework for benchmarking Class-out-of-distribution detection and its application to ImageNet

Ido Galil,Mohammed Dabbah,Ran El-Yaniv

When deployed for risk-sensitive tasks, deep neural networks must be able to detect instances with labels from outside the distribution for which they were trained.

In this paper we present a novel framework to benchmark the ability of image classifiers to detect class-out-of-distribution instances

(i.e., instances whose true labels do not appear in the training distribution) at various levels of detection difficulty.

We apply this technique to ImageNet, and benchmark 525 pretrained, publicly available, ImageNet-1k classifiers.

The code for generating a benchmark for any ImageNet-1k classifier, along with the benchmarks prepared for the above-mentioned 525 models is available at https://github.com/mdabbah/COOD_benchmarking.

The usefulness of the proposed framework and its advantage over alternative existing benchmarks is demonstrated by analyzing the results obtained for these models, which reveals numerous novel observations including:

(1) knowledge distillation consistently improves class-out-of-distribution (C-OOD) detection performance; (2) a subset of ViTs performs better C-OOD detection than any other model; (3) the language–-vision CLIP model achieves good zero-shot detection performance, with its best instance outperforming 96% of all other models evaluated; (4) accuracy and in-distribution ranking are positively correlated to C-OOD detection; and

(5) we compare various confidence functions for C-OOD detection.

Our companion paper, also published in ICLR 2023 (What Can We Learn From The Sel ective Prediction And Uncertainty Estimation Performance Of 523 Imagenet Classifiers), examines the uncertainty estimation performance (ranking, calibration, and selective prediction performance) of these classifiers in an in-distribution setting.

****

Data Poisoning Attacks Against Multimodal Encoders

Ziqing Yang,Xinlei He,Zheng Li,Michael Backes,Mathias Humbert,Pascal Berrang,Yang Zhang

Traditional machine learning (ML) models, e.g., image classifiers, usually rely on large-scale labeled datasets to achieve strong performance. However, such labeled datasets are often challenging and expensive to obtain. Also, the predefined categories limit the model's ability to generalize to other visual concepts as additional labeled data is required. On the contrary, the newly emerged multimodal model, which contains both visual and linguistic modalities, learns the concept of images from the raw text. It is a promising way to solve the above problems as it can use easy-to-collect image-text pairs to construct the training dataset and the raw texts contain almost unlimited categories according to their semantics. However, learning from a large-scale unlabeled dataset also exposes the model to the risk of potential poisoning attacks, whereby the adversary aims to perturb the model's training dataset to trigger malicious behaviors in it. Previous work mainly focuses on the visual modality. In this paper, we instead focus on answering two questions: (1) Is the linguistic modality also vulnerable to poisoning attacks? and (2) Which modality is most vulnerable? To answer the two questions, we conduct three types of poisoning attacks against CLIP, the most representative multimodal contrastive learning framework. Extensive evaluations on different datasets and model architectures show that all three attacks can perform well on the linguistic modality with only a relatively low poisoning rate and limited epochs. Also, we observe that the poisoning effect differs between different modalities, i.e., with lower MinRank in the visual modality and with higher Hit@K when K is small in the linguistic modality. To mitigate the attacks, we propose both pre-training and post-training defenses. We empirically show that both defenses can significantly reduce the attack performance while preserving the model's utility.

****

SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models

Ziyi Wu,Nikita Dvornik,Klaus Greff,Thomas Kipf,Animesh Garg

Understanding dynamics from visual observations is a challenging problem that requires disentangling individual objects from the scene and learning their interactions. While recent object-centric models can successfully decompose a scene into objects, modeling their dynamics effectively still remains a challenge. We address this problem by introducing SlotFormer -- a Transformer-based autoregressive model operating on learned object-centric representations. Given a video clip, our approach reasons over object features to model spatio-temporal relationships and predicts accurate future object states. In this paper, we successfully apply SlotFormer to perform video prediction on datasets with complex object interactions. Moreover, the unsupervised SlotFormer's dynamics model can be used to improve the performance on supervised downstream tasks, such as Visual Question Answering (VQA), and goal-conditioned planning. Compared to past works on dynamics modeling, our method achieves significantly better long-term synthesis of object dynamics, while retaining high quality visual generation. Besides, SlotFormer enables VQA models to reason about the future without object-level labels, even outperforming counterparts that use ground-truth annotations. Finally, we show its ability to serve as a world model for model-based planning, which is competitive with methods designed specifically for such tasks.

****

Simplifying Model-based RL: Learning Representations, Latent-space Models, and Policies with One Objective

Raj Ghugare,Homanga Bharadhwaj,Benjamin Eysenbach,Sergey Levine,Russ Salakhutdinov

While reinforcement learning (RL) methods that learn an internal model of the en vironment have the potential to be more sample efficient than their model-free counterparts, learning to model raw observations from high dimensional sensors can be challenging.

Prior work has addressed this challenge by learning low-dimensional representation of observations through auxiliary objectives, such as reconstruction or value prediction. However, the alignment between these auxiliary objectives and the RL objective is often unclear.

In this work, we propose a single objective which jointly optimizes a latent-space model and policy to achieve high returns while remaining self-consistent. This objective is a lower bound on expected returns. Unlike prior bounds for model-based RL on policy exploration or model guarantees, our bound is directly on the overall RL objective. We demonstrate that the resulting algorithm matches or improves the sample-efficiency of the best prior model-based and model-free RL methods. While sample efficient methods typically are computationally demanding, our method attains the performance of SAC in about 50\% less wall-clock time.

****

Tessellated Neural Networks: A Robust Defence against Adversarial Attacks

Chinmay Jain,Pabitra Mitra,Debasis Ganguly

Data-driven deep learning approaches for image classification are prone to adversarial attacks. An adversarial image which is sufficiently close (visually indistinguishable) from a true image of its representative class can often be misclassified to be a member of a different class. It is possible for attackers to exploit the high dimensionality of image representations, as learned by the neural models, to identify adversarial perturbations. To mitigate this problem, we propose a novel divide-and-conquer based approach of tessellating a base network architecture (e.g., a ResNet used in our experiments). The tessellated network learns the parameterized representations of each non-overlapping sub-region or tiles within an image, independently, and then learns how to combine these representations to finally estimate the class of the input image. We investigate two different modes of tessellation, namely periodic, comprised of regular square-shaped tiles, and aperiodic, comprised of rectangles of different dimensions. Experiments demonstrate that the tessellated extension of two standard deep neural models leads to a better defence against a number of standard adversarial attacks. We observed that the decrease in post-attack accuracy values relative to the accuracy of the uncompromised networks is smaller for our proposed tessellated approach.

****

Retrieval-based Controllable Molecule Generation

Zichao Wang,Weili Nie,Zhuoran Qiao,Chaowei Xiao,Richard Baraniuk,Anima Anandkumar

Generating new molecules with specified chemical and biological properties via generative models has emerged as a promising direction for drug discovery. However, existing methods require extensive training/fine-tuning with a large dataset, often unavailable in real-world generation tasks. In this work, we propose a new retrieval-based framework for controllable molecule generation. We use a small set of exemplar molecules, i.e., those that (partially) satisfy the design criteria, to steer the pre-trained generative model towards synthesizing molecules that satisfy the given design criteria. We design a retrieval mechanism that retrieves and fuses the exemplar molecules with the input molecule, which is trained by a new self-supervised objective that predicts the nearest neighbor of the input molecule. We also propose an iterative refinement process to dynamically update the generated molecules and retrieval database for better generalization. Our approach is agnostic to the choice of generative models and requires no task-specific fine-tuning. On various tasks ranging from simple design criteria to a challenging real-world scenario for designing lead compounds that bind to the SARS-CoV-2 main protease, we demonstrate our approach extrapolates well beyond the retrieval database, and achieves better performance and wider applicability than previous methods.

****

ELRT: Towards Efficient Low-Rank Training for Compact Neural Networks

Yang Sui,Miao Yin,Wanzhao Yang,Yu Gong,Jinqi Xiao,Huy Phan,Ding Ding,Xiaozhong Xu,Shan Liu,Zhenzhong Chen,Bo Yuan

Low-rank compression, a popular model compression technique that produces compact convolutional neural networks (CNNs) with low rankness, has been well studied in the literature. On the other hand, low-rank training, as an alternative way to train low-rank CNNs from scratch, is little exploited yet. Unlike low-rank compression, low-rank training does not need pre-trained full-rank models and the entire training phase is always performed on the low-rank structure, bringing attractive benefits for practical applications. However, the existing low-rank training solutions are still very limited and do not demonstrate their effectiveness for training modern low-rank CNN models in the large-scale dataset from scratch. In this paper, we perform a systematic investigation on low-rank CNN training. By identifying the proper low-rank format and performance-improving strategy, we propose ELRT, an efficient low-rank training solution for high-accuracy high-compactness low-rank CNN models. Our extensive evaluation results for training various CNNs on different datasets demonstrate the effectiveness of ELRT.

****

InfoOT: Information Maximizing Optimal Transport

Ching-Yao Chuang,Stefanie Jegelka,David Alvarez-Melis

Optimal transport aligns samples across distributions by minimizing the transportation cost between them, e.g., the geometric distances. Yet, it ignores coherence structure in the data such as clusters, does not handle outliers well, and cannot integrate new data points. To address these drawbacks, we propose InfoOT, an information-theoretic extension of optimal transport that maximizes the mutual information between domains while minimizing geometric distances. The resulting objective can still be formulated as a (generalized) optimal transport problem, and can be efficiently solved by projected gradient descent. This formulation yields a new projection method that is robust to outliers and generalizes to unseen samples. Empirically, InfoOT improves the quality of alignments across benchmarks in domain adaptation, cross-domain retrieval, and single-cell alignment.

****

Posterior Sampling Model-based Policy Optimization under Approximate Inference

Chaoqi Wang,Yuxin Chen,Kevin Patrick Murphy

Model-based reinforcement learning algorithms (MBRL) hold tremendous promise for improving the sample efficiency in online RL. However, many existing popular MBRL algorithms cannot deal with exploration and exploitation properly. Posterior sampling reinforcement learning (PSRL) serves as a promising approach for automatically trading off the exploration and exploitation, but the theoretical guarantees only hold under exact inference. In this paper, we show that adopting the same methodology as in exact PSRL can be fairly suboptimal under approximate inference. Motivated by the analysis, we propose an improved factorization for the posterior distribution of polices by removing the conditional independence between the policy and data given the model. By adopting such a posterior factorization, we further propose a general algorithmic framework for PSRL under approximate inference and a practical instantiation of it. Empirically, our algorithm can surpass the baseline methods by a significant margin on both dense rewards and sparse rewards tasks from DM control suite, OpenAI Gym and Metaworld benchmarks.

****

Causal discovery from conditionally stationary time series

Carles Balsells Rodas,Ruibo Tu,Hedvig Kjellstrom,Yingzhen Li

Causal discovery, i.e., inferring underlying causal relationships from observational data, has been shown to be highly challenging for AI systems. In time series modeling context, traditional causal discovery methods mainly consider constrained scenarios with fully observed variables and/or data from stationary time-series. We develop a causal discovery approach to handle a wide class of non-stationary time-series that are conditionally stationary, where the non-stationary behaviour is modeled as stationarity conditioned on a set of (possibly hidden) state variables. Named state-dependent causal inference (SDCI), our approach is able to recover the underlying causal dependencies, provably with fully-observed st ates and empirically with hidden states. The latter is confirmed by experiments on synthetic linear system and nonlinear particle interaction data, where SDCI achieves superior performance over baseline causal discovery methods. Improved results over non-causal RNNs on modeling NBA player movements demonstrate the potential of our method and motivate the use causality-driven methods for forecasting.

****

Learning for Edge-Weighted Online Bipartite Matching with Robustness Guarantees

Pengfei Li,Jianyi Yang,Shaolei Ren

Many real-world problems, such as online ad display, can be formulated as online bipartite matching. The crucial challenge lies in the nature of sequentially-revealed online item information, based on which we make irreversible matching decisions at each step. While numerous expert online algorithms have been proposed with bounded worst-case competitive ratios, they may not offer satisfactory performance in average cases. On the other hand, reinforcement learning (RL) has been applied to improve the average performance, but they lack robustness and can perform arbitrarily badly. In this paper, we propose a novel RL-based approach to edge-weighted online bipartite matching with robustness guarantees (LOMAR), achieving both good average-case and good worst-case performance. The key novelty of LOMAR is a new online switching operation which, based on a judiciously-designed condition to hedge against future uncertainties, decides whether to follow the expert's decision or the RL decision for each online item arrival. We prove that for any $\rho \in [0,1]$, LOMAR is ρ-competitive against any given expert online algorithm. To improve the average performance, we train the RL policy by explicitly considering the online switching operation. Finally, we run empirical experiments to demonstrate the advantages of LOMAR compared to existing baselines.

****

Tangential Wasserstein Projections

Florian Gunsilius,Meng Hsuan Hsieh,Myung Jin Lee

We develop a notion of projections between sets of probability measures using the geometric properties of the 2-Wasserstein space. It is designed for general multivariate probability measures, is computationally efficient to implement, and provides a unique solution in regular settings. The idea is to work on regular tangent cones of the Wasserstein space using generalized geodesics. Its structure and computational properties make the method applicable in a variety of settings, from causal inference to the analysis of object data. An application to estimating causal effects yields a generalization of the notion of synthetic controls for systems with general heterogeneity described via multivariate probability measures, as well as a way to estimate optimal weights jointly over all time periods.

****

Data Drift Correction via Time-varying Importance Weight Estimator

Rasool Fakoor,Jonas Mueller,Zachary Chase Lipton,Pratik Chaudhari,Alex Smola

Real-world deployment of machine learning models is challenging when data evolves over time. And data does evolve 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 novel time-varying importance weight estimator that can detect gradual shifts in the distribution of data. Such an importance weight estimator allows the training method to selectively sample past data---not just similar data from the past like a standard importance weight estimator would but also data that evolved in a similar fashion in the past. Our time-varying importance weight is quite general. We demonstrate different ways of implementing it that exploit some known structure in the evolution of data. We demonstrate and evaluate this approach on a variety of problems ranging from supervised learning tasks (multiple image classification datasets) where the data undergoes a sequence of gradual shifts of our design to reinforcement learning tasks (robotic manipulation and continuous control) where data undergoes a shift organically as the policy or the task changes.

****

Analytical Composition of Differential Privacy via the Edgeworth Accountant
Hua Wang,Sheng Gao,Huanyu Zhang,Milan Shen,Weijie J Su

Many modern machine learning algorithms are composed of simple private algorithms; thus, an increasingly important problem is to efficiently compute the overall privacy loss under composition. In this study, we introduce the Edgeworth Accountant, an analytical approach to composing differential privacy guarantees of private algorithms. The Edgeworth Accountant starts by losslessly tracking the privacy loss under composition using the f-differential privacy framework, which allows us to express the privacy guarantees using privacy-loss log-likelihood ratios (PLLRs). As the name suggests, this accountant next uses the Edgeworth expansion to the upper and lower bounds the probability distribution of the sum of the PLLRs. Moreover, by relying on a technique for approximating complex distributions using simple ones, we demonstrate that the Edgeworth Accountant can be applied to the composition of any noise-addition mechanism. Owing to certain appealing features of the Edgeworth expansion, the (ϵ, δ)-differential privacy bounds offered by this accountant are non-asymptotic, with essentially no extra computational cost, as opposed to the prior approaches, wherein the running times increase with the number of compositions. Finally, we demonstrate that our upper and lower (ϵ, δ)-differential privacy bounds are tight in federated analytics and certain regimes of training private deep learning models.

****

Policy-Induced Self-Supervision Improves Representation Finetuning in Visual RL
Séb Arnold,Fei Sha

We study how to transfer representations pretrained on source tasks to target tasks in visual percept based RL. We analyze two popular approaches: freezing or finetuning the pretrained representations. Empirical studies on a set of popular tasks reveal several properties of pretrained representations. First, finetuning is required even when pretrained representations perfectly capture the information required to solve the target task. Second, finetuned representations improve learnability and are more robust to noise.

Third, pretrained bottom layers are task-agnostic and readily transferable to new tasks, while top layers encode task-specific information and require adaptation. Building on these insights, we propose a self-supervised objective that \emph{clusters representations according to the policy they induce}, as opposed to traditional representation similarity measures which are policy-agnostic (\eg Euclidean norm, cosine similarity). Together with freezing the bottom layers, this objective results in significantly better representation than frozen, finetuned, and self-supervised alternatives on a wide range of benchmarks.

****

Deep Generative Symbolic Regression
Samuel Holt,Zhaozhi Qian,Mihaela van der Schaar

Symbolic regression (SR) aims to discover concise closed-form mathematical equations from data, a task fundamental to scientific discovery. However, the problem is highly challenging because closed-form equations lie in a complex combinatorial search space. Existing methods, ranging from heuristic search to reinforcement learning, fail to scale with the number of input variables. We make the observation that closed-form equations often have structural characteristics and invariances (e.g. the commutative law) that could be further exploited to build more effective symbolic regression solutions. Motivated by this observation, our key contribution is to leverage pre-trained deep generative models to capture the intrinsic regularities of equations, thereby providing a solid foundation for subsequent optimization steps. We show that our novel formalism unifies several prominent approaches of symbolic regression and offers a new perspective to justify and improve on the previous ad hoc designs, such as the usage of cross-entropy loss during pre-training. Specifically, we propose an instantiation of our framework, Deep Generative Symbolic Regression (DGSR). In our experiments, we show that DGSR achieves a higher recovery rate of true equations in the setting of a larger number of input variables, and it is more computationally efficient at infe

rence time than state-of-the-art RL symbolic regression solutions.
****
What Can we Learn From The Selective Prediction And Uncertainty Estimation Performance Of 523 Imagenet Classifiers?
Ido Galil,Mohammed Dabbah,Ran El-Yaniv
When deployed for risk-sensitive tasks, deep neural networks must include an uncertainty estimation mechanism.
Here we examine the relationship between deep architectures and their respective training regimes, with their corresponding selective prediction and uncertainty estimation performance. We consider some of the most popular estimation performance metrics previously proposed including AUROC, ECE, AURC as well as coverage for selective accuracy constraint.
We present a novel and comprehensive study of selective prediction and the uncertainty estimation performance of 523 existing pretrained deep ImageNet classifiers that are available in popular repositories.
We identify numerous and previously unknown factors that affect uncertainty estimation and examine the relationships between the different metrics. We find that distillation-based training regimes consistently yield better uncertainty estimations than other training schemes such as vanilla training, pretraining on a larger dataset and adversarial training.
Moreover, we find a subset of ViT models that outperform any other models in terms of uncertainty estimation performance.
For example, we discovered an unprecedented 99% top-1 selective accuracy on ImageNet at 47% coverage
(and 95% top-1 accuracy at 80%) for a ViT model, whereas a competing EfficientNet-V2-XL cannot obtain these accuracy constraints at any level of coverage.
Our companion paper, also published in ICLR 2023 (A framework for benchmarking class-out-of-distribution detection and its application to ImageNet), examines the performance of these classifiers in a class-out-of-distribution setting.
****
Solving and Learning non-Markovian Stochastic Control problems in continuous-time with Neural RDEs
Melker Höglund,Emilio Ferrucci,Camilo Hernández,Aitor Muguruza Gonzalez,Cristopher Salvi,Leandro Sánchez-Betancourt,Yufei Zhang
We propose a novel framework for solving continuous-time, non-Markovian stochastic control problems with the use of neural rough differential equations (Neural RDEs). By parameterising the control process as the solution of a Neural RDE driven by the state process, we show that the control-state joint dynamics are governed by an uncontrolled RDE with structured vector fields, allowing for efficient trajectories simulation, Monte-Carlo estimation of the value function and backpropagation. To deal with input paths of infinite 1-variation, we refine the existing universal approximation result to a probabilistic density result for Neural RDEs driven by random rough paths. Experiments on various non-Markovian problems indicate how the proposed framework is time-resolution-invariant and capable of learning optimal solutions with higher accuracy than traditional RNN-based approaches. Finally, we discuss possible extensions of this framework to the setting of non-Markovian continuous-time reinforcement learning and provide promising empirical evidence in this direction.
****
Spatio-temporal Self-Attention for Egocentric 3D Pose Estimation
Jinman Park,Kimathi Kaai,Saad Hossain,Norikatsu Sumi,Sirisha Rambhatla,Paul W. Fieguth
Vision-based ego-centric 3D human pose estimation (ego-HPE) is essential to support critical applications of xR-technologies. However, severe self-occlusions and strong distortion introduced by the fish-eye view from the head mounted camera, make ego-HPE extremely challenging. While current state-of-the-art (SOTA) methods try to address the distortion, they still suffer from large errors in the most critical joints (such as hands) due to self-occlusions. To this end, we propose a spatio-temporal transformer model that can attend to semantically rich feature maps obtained from popular convolutional backbones. Leveraging the complex s

patio-temporal information encoded in ego-centric videos, we design a spatial concept called feature map tokens (FMT) which can attend to all the other spatial units in our spatio-temporal feature maps. Powered by this FMT-based transformer, we build Egocentric Spatio-Temporal Self-Attention Network (Ego-STAN), which uses heatmap-based representations and spatio-temporal attention specialized to address distortions and self-occlusions in ego-HPE.

Our quantitative evaluation on the contemporary sequential xR-EgoPose dataset, achieves a 38.2% improvement on the highest error joints against the SOTA ego-HPE model, while accomplishing a 22% decrease in the number of parameters. Finally, we also demonstrate the generalization capabilities of our model to real-world HPE tasks beyond ego-views.

****

RNAS-CL: Robust Neural Architecture Search by Cross-Layer Knowledge Distillation

Utkarsh Nath,Yancheng Wang,Yingzhen Yang

Deep Neural Networks are vulnerable to adversarial attacks. Neural Architecture Search (NAS), one of the driving tools of deep neural networks, demonstrates superior performance in prediction accuracy in various machine learning applications. However, it is unclear how it performs against adversarial attacks. Given the presence of a robust teacher, it would be interesting to investigate if NAS would produce robust neural architecture by inheriting robustness from the teacher. In this paper, we propose Robust Neural Architecture Search by Cross-Layer Knowledge Distillation (RNAS-CL), a novel NAS algorithm that improves the robustness of NAS by learning from a robust teacher through cross-layer knowledge distillation. Unlike previous knowledge distillation methods that encourage close student/teacher output only in the last layer, RNAS-CL automatically searches for the best teacher layer to supervise each student layer. Experimental result evidences the effectiveness of RNAS-CL and shows that RNAS-CL produces small and robust neural architecture.

****

Multi-Agent Policy Transfer via Task Relationship Modeling

Rong-Jun Qin,Feng Chen,Tonghan Wang,Lei Yuan,Xiaoran Wu,Yipeng Kang,Zongzhang Zhang,Chongjie Zhang,Yang Yu

Team adaptation to new cooperative tasks is a hallmark of human intelligence, which has yet to be fully realized in learning agents. Previous works on multi-agent transfer learning accommodate teams of different sizes, but heavily rely on the generalization ability of neural networks for adapting to unseen tasks. We posit that the relationship among tasks provides the key information for policy adaptation. To utilize such relationship for efficient transfer, we try to discover and exploit the knowledge among tasks from different teams, propose to learn effect-based task representations as a common latent space among tasks, and use it to build an alternatively fixed training scheme. We demonstrate that the task representation can capture the relationship among teams and generalize to unseen tasks. As a result, the proposed method can help transfer learned cooperation knowledge to new tasks after training on a few source tasks, and the learned transferred policies can also help solve tasks that are hard to learn from scratch.

****

Predictor-corrector algorithms for stochastic optimization under gradual distribution shift

Subha Maity,Debarghya Mukherjee,Moulinath Banerjee,Yuekai Sun

Time-varying stochastic optimization problems frequently arise in machine learning practice (e.g., gradual domain shift, object tracking, strategic classification). Often, the underlying process that drives the distribution shift is continuous in nature. We exploit this underlying continuity by developing predictor-corrector algorithms for time-varying stochastic optimization that anticipates changes in the underlying data generating process through a predictor-corrector term in the update rule. The key challenge is the estimation of the predictor-corrector term; a naive approach based on sample-average approximation may lead to non-convergence. We develop a general moving-average based method to estimate the predictor-corrector term and provide error bounds for the iterates, both in presence of pure and noisy access to the queries from the relevant derivatives of th e loss function. Furthermore, we show (theoretically and empirically in several examples) that our method outperforms non-predictor corrector methods that do not anticipate changes in the data generating process.

****

AIM: Adapting Image Models for Efficient Video Action Recognition

Taojiannan Yang,Yi Zhu,Yusheng Xie,Aston Zhang,Chen Chen,Mu Li

Recent vision transformer based video models mostly follow the ``image pre-training then finetuning" paradigm and have achieved great success on multiple video benchmarks. However, fully finetuning such a video model could be computationally expensive and unnecessary, given the pre-trained image transformer models have demonstrated exceptional transferability. In this work, we propose a novel method to Adapt pre-trained Image Models (AIM) for efficient video understanding. By freezing the pre-trained image model and adding a few lightweight Adapters, we introduce spatial adaptation, temporal adaptation and joint adaptation to gradually equip an image model with spatiotemporal reasoning capability. We show that our proposed AIM can achieve competitive or even better performance than prior arts with substantially fewer tunable parameters on four video action recognition benchmarks. Thanks to its simplicity, our method is also generally applicable to different image pre-trained models, which has the potential to leverage more powerful image foundation models in the future. The project webpage is https://adapt-image-models.github.io/.

****

Impossibly Good Experts and How to Follow Them

Aaron Walsman,Muru Zhang,Sanjiban Choudhury,Dieter Fox,Ali Farhadi

We consider the sequential decision making problem of learning from an expert that has access to more information than the learner. For many problems this extra information will enable the expert to achieve greater long term reward than any policy without this privileged information access. We call these experts ``Impossibly Good'' because no learning algorithm will be able to reproduce their behavior. However, in these settings it is reasonable to attempt to recover the best policy possible given the agent's restricted access to information. We provide a set of necessary criteria on the expert that will allow a learner to recover the optimal policy in the reduced information space from the expert's advice alone. We also provide a new approach called Elf Distillation (Explorer Learning from Follower) that can be used in cases where these criteria are not met and environmental rewards must be taken into account. We show that this algorithm performs better than a variety of strong baselines on a challenging suite of Minigrid and Vizdoom environments.

****

On Convergence of Average-Reward Off-Policy Control Algorithms in Weakly-Communicating MDPs

Yi Wan,Richard S. Sutton

We show two average-reward off-policy control algorithms, Differential Q Learning (Wan, Naik, \& Sutton 2021a) and RVI Q Learning (Abounadi Bertsekas \& Borkar 2001), converge in weakly-communicating MDPs. Weakly-communicating MDPs are the most general class of MDPs that a learning algorithm with a single stream of experience can guarantee obtaining a policy achieving optimal reward rate. The original convergence proofs of the two algorithms require that all optimal policies induce unichains, which is not necessarily true for weakly-communicating MDPs. To the best of our knowledge, our results are the first showing average-reward off-policy control algorithms converge in weakly-communicating MDPs. As a direct extension, we show that average-reward options algorithms introduced by (Wan, Naik, \& Sutton 2021b) converge if the Semi-MDP induced by options is weakly-communicating.

****

Distributionally Robust Post-hoc Classifiers under Prior Shifts

Jiaheng Wei,Harikrishna Narasimhan,Ehsan Amid,Wen-Sheng Chu,Yang Liu,Abhishek Kumar

The generalization ability of machine learning models degrades significantly when the test distribution shifts away from the training distribution. We investiga te the problem of training models that are robust to shifts caused by changes in the distribution of class-priors or group-priors. The presence of skewed training priors can often lead to the models overfitting to spurious features. Unlike existing methods, which optimize for either the worst or the average performance over classes or groups, our work is motivated by the need for finer control over the robustness properties of the model. We present an extremely lightweight post-hoc approach that performs scaling adjustments to predictions from a pre-trained model, with the goal of minimizing a distributionally robust loss around a chosen target distribution. These adjustments are computed by solving a constrained optimization problem on a validation set and applied to the model during test time. Our constrained optimization objective is inspired from a natural notion of robustness to controlled distribution shifts. Our method comes with provable guarantees and empirically makes a strong case for distributional robust post-hoc classifiers. An empirical implementation is available at https://github.com/weijiaheng/Drops.

****

Transformer Meets Boundary Value Inverse Problems

Ruchi Guo,Shuhao Cao,Long Chen

A Transformer-based deep direct sampling method is proposed for electrical impedance tomography, a well-known severely ill-posed nonlinear boundary value inverse problem. A real-time reconstruction is achieved by evaluating the learned inverse operator between carefully designed data and the reconstructed images. An effort is made to give a specific example to a fundamental question: whether and how one can benefit from the theoretical structure of a mathematical problem to develop task-oriented and structure-conforming deep neural networks? Specifically, inspired by direct sampling methods for inverse problems, the 1D boundary data in different frequencies are preprocessed by a partial differential equation-based feature map to yield 2D harmonic extensions as different input channels. Then, by introducing learnable non-local kernels, the direct sampling is recast to a modified attention mechanism. The new method achieves superior accuracy over its predecessors and contemporary operator learners and shows robustness to noises in benchmarks.

This research shall strengthen the insights that, despite being invented for natural language processing tasks, the attention mechanism offers great flexibility to be modified in conformity with the a priori mathematical knowledge, which ultimately leads to the design of more physics-compatible neural architectures.

****

Diagnosing and exploiting the computational demands of videos games for deep reinforcement learning

Lakshmi Narasimhan Govindarajan,Rex G Liu,Drew Linsley,Alekh Karkada Ashok,Max Reuter,Michael Frank,Thomas Serre

Humans learn by interacting with their environments and perceiving the outcomes of their actions. A landmark in artificial intelligence has been the development of deep reinforcement learning (dRL) algorithms capable of doing the same in video games, on par with or better than humans. However, it remains unclear whether the successes of dRL models reflect advances in visual representation learning, the effectiveness of reinforcement learning algorithms at discovering better policies, or both. To address this question, we introduce the Learning Challenge Diagnosticator (LCD), a tool that separately measures the perceptual and reinforcement learning demands of a task. We use LCD to discover a novel taxonomy of challenges in the Procgen benchmark, and demonstrate that these predictions are both highly reliable and can instruct algorithmic development. More broadly, the LCD reveals multiple failure cases that can occur when optimizing dRL algorithms over entire video game benchmarks like Procgen, and provides a pathway towards more efficient progress.

****

NeuralPCG: Learning Preconditioner for Solving Partial Differential Equations with Graph Neural Network

Yichen Li,Tao Du,Peter Yichen Chen,Wojciech Matusik

Fast and accurate partial differential equation (PDE) solvers empower scientific and engineering research. Classic numerical solvers provide unparalleled accuracy but often require extensive computation time. Machine learning solvers are significantly faster but lack convergence and accuracy guarantees. We present Neural-Network-Preconditioned Conjugate Gradient, or NeuralPCG, a novel linear second-order PDE solver that combines the benefits of classic iterative solvers and machine learning approaches. Our key observation is that both neural-network PDE solvers and classic preconditioners excel at obtaining fast but inexact solutions. NeuralPCG proposes to use neural network models to \emph{precondition} PDE systems in classic iterative solvers. Compared with neural-network PDE solvers, NeuralPCG achieves converging and accurate solutions (e.g.,1e-12 precision) by construction. Compared with classic solvers, NeuralPCG is faster via data-driven preconditioners. We demonstrate the efficacy and generalizability of NeuralPCG by conducting extensive experiments on various 2D and 3D linear second-order PDEs.
****

Learning Dynamic Query Combinations for Transformer-based Object Detection and Segmentation
Yiming Cui,Linjie Yang,Haichao Yu

Transformer-based detection and segmentation methods use a list of learned detection queries to retrieve information from the transformer network and learn to predict the location and category of one specific object from each query. We empirically find that random convex combinations of the learned queries are still good queries for the corresponding models. We then propose to learn a convex combination with dynamic coefficients based on the high-level semantics of the image. The generated dynamic queries better capture the prior of object locations and categories in the different images. Equipped with our dynamic queries, a wide range of DETR-based models achieve consistent and superior performance across multiple tasks (object detection, instance segmentation, panoptic segmentation) and on different benchmarks (MS COCO, CityScapes, YoutubeVIS).
****

Cross-Quality Few-Shot Transfer for Alloy Yield Strength Prediction: A New Material Science Benchmark and An Integrated Optimization Framework
Xuxi Chen,Tianlong Chen,Everardo Yeriel Olivares,Kate Elder,Scott K. McCall,Aurelien Pierre Philippe Perron,Joseph T. McKeown,Bhavya Kailkhura,Zhangyang Wang,Brian Gallagher

Discovering high-entropy alloys (HEAs) with high yield strength is an important yet challenging task in material science. However, the yield strength can only be accurately measured by very expensive and time-consuming real-world experiments, hence cannot be acquired at scale. Learning-based methods could facilitate the discovery process, but the lack of a comprehensive dataset on HEA yield strength has created barriers. We present X-Yield, a large-scale material science benchmark with 240 experimentally measured ("high-quality") and over 100K simulated (imperfect or "low-quality") HEA yield strength annotations. Due to the scarcity of experimental annotations and the quality gap in imperfectly simulated data, existing transfer learning methods cannot generalize well on our dataset. We address this cross-quality few-shot transfer problem by leveraging model sparsification "twice" --- as a noise-robust feature learning regularizer at the pre-training stage, and as a data-efficient learning regularizer at the few-shot transfer stage. While the workflow already performs decently with ad-hoc sparsity patterns tuned independently for either stage, we take a step further by proposing a bi-level optimization framework termed Bi-RPT, that jointly learns optimal masks and automatically allocates sparsity levels for both stages. The optimization problem is solved efficiently using gradient unrolling, which is seamlessly integrated with the training process. The effectiveness of Bi-RPT is validated through extensive experiments on our new challenging X-Yield dataset, alongside other synthesized testbeds. Specifically, we achieve an 8.9~19.8% reduction in terms of the test mean squared error and 0.98~1.53% in terms of test accuracy, merely using 5-10% of the experimental data. Codes and sample data are in the supplement.
****

Robust Reinforcement Learning with Distributional Risk-averse formulation

Pierre Clavier,Stephanie Allassonniere,Erwan Le Pennec

The purpose of robust reinforcement learning is to make predictions more robust to changes in the dynamics or rewards of the system. This problem is particularly important when dynamics and rewards of the environment are estimated from the data. However, without constraints, this problem is intractable. In this paper, we approximate the Robust Reinforcement Learning constrained with a f-divergence using an approximate Risk-Averse formulation. We show that the classical Reinforcement Learning formulation can be robustified using a standard deviation penalization of the objective. Two algorithms based on Distributional Reinforcement Learning, one for discrete and one for continuous action spaces, are proposed and tested in a classical Gym environment to demonstrate the robustness of the algorithms.

****

Unicom: Universal and Compact Representation Learning for Image Retrieval

Xiang An,Jiankang Deng,Kaicheng Yang,Jaiwei Li,Ziyong Feng,Jia Guo,Jing Yang,Tongliang Liu

Modern image retrieval methods typically rely on fine-tuning pre-trained encoders to extract image-level descriptors.

However, the most widely used models are pre-trained on ImageNet-1K with limited classes. The pre-trained feature representation is therefore not universal enough to generalize well to the diverse open-world classes.

In this paper, we first cluster the large-scale \laion{} into one million pseudo classes based on the joint textual and visual features extracted by the CLIP model. Due to the confusion of label granularity, the automatically clustered dataset inevitably contains heavy inter-class conflict. To alleviate such conflict, we randomly select partial inter-class prototypes to construct the margin-based softmax loss. To further enhance the low-dimensional feature representation, we randomly select partial feature dimensions when calculating the similarities between embeddings and class-wise prototypes. The dual random partial selections are with respect to the class dimension and the feature dimension of the prototype matrix, making the classification conflict-robust and the feature embedding compact. Our method significantly outperforms state-of-the-art unsupervised and supervised image retrieval approaches on multiple benchmarks. The code and pre-trained models are released to facilitate future research \url{https://github.com/deepglint/unicom}.

****

The Reward Hypothesis is False

Joar Max Viktor Skalse,Alessandro Abate

The reward hypothesis is the hypothesis that "all of what we mean by goals and purposes can be well thought of as the maximisation of the expected value of the cumulative sum of a received scalar signal". In this paper, we will argue that this hypothesis is false. We will look at three natural classes of reinforcement learning tasks (multi-objective reinforcement learning, risk-averse reinforcement learning, and modal reinforcement learning), and then prove mathematically that these tasks cannot be expressed using any scalar, Markovian reward function. We thus disprove the reward hypothesis by providing many examples of tasks which are both natural and intuitive to describe, but which are nonetheless impossible to express using reward functions. In the process, we provide necessary and sufficient conditions for when a multi-objective reinforcement learning problem can be reduced to ordinary, scalar reward reinforcement learning. We also call attention to a new class of reinforcement learning problems (namely those we call "modal" problems), which have so far not been given any systematic treatment in the reinforcement learning literature.

****

Convergence of Generative Deep Linear Networks Trained with Bures-Wasserstein Loss

Pierre Bréchet,Katerina Papagiannouli,Jing An,Guido Montufar

We consider a deep matrix factorization model of covariance matrices trained with the Bures-Wasserstein distance. While recent works have made important advances in the study of the optimization problem for overparametrized low-rank matrix approximation, much emphasis has been placed on discriminative settings and the square loss. In contrast, our model considers another interesting type of loss and connects with the generative setting. We characterize the critical points and minimizers of the Bures-Wasserstein distance over the space of rank-bounded matrices. For low-rank matrices the Hessian of this loss can blow up, which creates challenges to analyze convergence of optimizaton methods. We establish convergence results for gradient flow using a smooth perturbative version of the loss and convergence results for finite step size gradient descent under certain assumptions on the initial weights

****

Diffusion Probabilistic Fields

Peiye Zhuang,Samira Abnar,Jiatao Gu,Alex Schwing,Joshua M. Susskind,Miguel Ángel Bautista

Diffusion probabilistic models have quickly become a major approach for generative modeling of images, 3D geometry, video and other domains. However, to adapt diffusion generative modeling to these domains the denoising network needs to be carefully designed for each domain independently, oftentimes under the assumption that data lives in a Euclidean grid. In this paper we introduce Diffusion Probabilistic Fields (DPF), a diffusion model that can learn distributions over continuous functions defined over metric spaces, commonly known as fields. We extend the formulation of diffusion probabilistic models to deal with this field parametrization in an explicit way, enabling us to define an end-to-end learning algorithm that side-steps the requirement of representing fields with latent vectors as in previous approaches (Dupont et al., 2022a; Du et al., 2021). We empirically show that, while using the same denoising network, DPF effectively deals with different modalities like 2D images and 3D geometry, in addition to modeling distributions over fields defined on non-Euclidean metric spaces.

****

Improving Information Retention in Large Scale Online Continual Learning

zhipeng cai,Vladlen Koltun,Ozan Sener

Given a stream of data sampled from non-stationary distributions, online continual learning (OCL) aims to adapt efficiently to new data while retaining existing knowledge. The typical approach to address information retention (the ability to retain previous knowledge) is keeping a replay buffer of a fixed size and computing gradients using a mixture of new data and the replay buffer. Surprisingly, the recent work (Cai et al., 2021) suggests that information retention remains a problem in large scale OCL even when the replay buffer is unlimited, \emph{i.e.}, the gradients are computed using all past data. This paper focuses on this peculiarity to understand and address information retention. To pinpoint the source of this problem, we theoretically show that, given limited computation budgets at each time step, even without strict storage limit, naively applying SGD with constant or constantly decreasing learning rates fail to optimize information retention in the long term. We propose using a moving average family of methods to improve optimization for non-stationary objectives. Specifically, we design an adaptive moving average (AMA) optimizer and a moving-average-based learning rate schedule (MALR). We demonstrate the effectiveness of AMA+MALR on large scale benchmarks, including Continual Localization (CLOC), Google Landmarks and ImageNet. Code
will be released upon publication.

****

Landscape Learning for Neural Network Inversion

Ruoshi Liu,Chengzhi Mao,Purva Tendulkar,Hao Wang,Carl Vondrick

Many machine learning methods operate by inverting a neural network at inference time, which has become a popular technique for solving inverse problems in computer vision, robotics, and graphics. However, these methods often involve gradient descent through a highly non-convex loss landscape, causing the optimization process to be unstable and slow. We introduce a method that learns a loss landscape where gradient descent is efficient, bringing massive improvement and acceleration to the inversion process. We demonstrate this advantage on a number of methods for both generative and discriminative tasks, including GAN inversion, adv

ersarial defense, and 3D human pose reconstruction.
****

Stochastic Multi-Person 3D Motion Forecasting

Sirui Xu,Yu-Xiong Wang,Liangyan Gui

This paper aims to deal with the ignored real-world complexity in prior work on human motion forecasting, emphasizing the social properties of multi-person motion, the diversity of motion and social interactions, and the complexity of articulated motion. To this end, we introduce a novel task of stochastic multi-person 3D motion forecasting. We propose a dual-level generative modeling framework that separately models independent individual motion at the local level and social interactions at the global level. Notably, this dual-level modeling mechanism can be achieved within a shared generative model, through introducing learnable latent codes that represent intents of future motion and switching the codes' modes of operation at different levels. Our framework is general, and we instantiate it with various multi-person forecasting models. Extensive experiments on CMU-Mocap, MuPoTS-3D, and SoMoF benchmarks show that our approach produces diverse and accurate multi-person predictions, significantly outperforming the state of the art.
****

ON INJECTING NOISE DURING INFERENCE

Milad Khademi Nori,Yiqun GE,IL MIN KIM

We study activation noise in a generative energy-based modeling setting during training for the purpose of regularization. We prove that activation noise is a general form of dropout. Then, we analyze the role of activation noise at inference time and demonstrate it to be utilizing sampling. Thanks to the activation noise we observe about 200% improvement in performance (classification accuracy). Later, we not only discover, but also prove that the best performance is achieved when the activation noise follows the same distribution both during training and inference. To explicate this phenomenon, we provide theoretical results that illuminate the roles of activation noise during training, inference, and their mutual influence on the performance. To further confirm our theoretical results, we conduct experiments for five datasets and seven distributions of activation noise.
****

LEARNING THE SPECTROGRAM TEMPORAL RESOLUTION FOR AUDIO CLASSIFICATION

Haohe Liu,Xubo Liu,Qiuqiang Kong,Wenwu Wang,Mark D Plumbley

The audio spectrogram is a time-frequency representation that has been widely used for audio classification. The temporal resolution of a spectrogram depends on hop size. Previous works generally assume the hop size should be a constant value such as ten milliseconds. However, a fixed hop size or resolution is not always optimal for different types of sound. This paper proposes a novel method, DiffRes, that enables differentiable temporal resolution learning to improve the performance of audio classification models. Given a spectrogram calculated with a fixed hop size, DiffRes merges non-essential time frames while preserving important frames. DiffRes acts as a "drop-in" module between an audio spectrogram and a classifier, and can be end-to-end optimized. We evaluate DiffRes on the mel-spectrogram, followed by state-of-the-art classifier backbones, and apply it to five different subtasks. Compared with using the fixed-resolution mel-spectrogram, the DiffRes-based method can achieve the same or better classification accuracy with at least 25% fewer temporal dimensions on the feature level, which alleviates the computational cost at the same time. Starting from a high-temporal-resolution spectrogram such as one-millisecond hop size, we show that DiffRes can improve classification accuracy with the same computational complexity.
****

Beyond calibration: estimating the grouping loss of modern neural networks

Alexandre Perez-Lebel,Marine Le Morvan,Gael Varoquaux

The ability to ensure that a classifier gives reliable confidence scores is essential to ensure informed decision-making. To this end, recent work has focused on miscalibration, i.e., the over or under confidence of model scores. Yet calibration is not enough: even a perfectly calibrated classifier with the best possib le accuracy can have confidence scores that are far from the true posterior probabilities. This is due to the grouping loss, created by samples with the same confidence scores but different true posterior probabilities. Proper scoring rule theory shows that given the calibration loss, the missing piece to characterize individual errors is the grouping loss. While there are many estimators of the calibration loss, none exists for the grouping loss in standard settings. Here, we propose an estimator to approximate the grouping loss. We show that modern neural network architectures in vision and NLP exhibit grouping loss, notably in distribution shifts settings, which highlights the importance of pre-production validation.

****

Hybrid RL: Using both offline and online data can make RL efficient

Yuda Song,Yifei Zhou,Ayush Sekhari,Drew Bagnell,Akshay Krishnamurthy,Wen Sun

We consider a hybrid reinforcement learning setting (Hybrid RL), in which an agent has access to an offline dataset and the ability to collect experience via real-world online interaction. The framework mitigates the challenges that arise in both pure offline and online RL settings, allowing for the design of simple and highly effective algorithms, in both theory and practice. We demonstrate these advantages by adapting the classical Q learning/iteration algorithm to the hybrid setting, which we call Hybrid Q-Learning or Hy-Q. In our theoretical results, we prove that the algorithm is both computationally and statistically efficient whenever the offline dataset supports a high-quality policy and the environment has bounded bilinear rank. Notably, we require no assumptions on the coverage provided by the initial distribution, in contrast with guarantees for policy gradient/iteration methods. In our experimental results, we show that Hy-Q with neural network function approximation outperforms state-of-the-art online, offline, and hybrid RL baselines on challenging benchmarks, including Montezuma’s Revenge.

****

Spotting Expressivity Bottlenecks and Fixing Them Optimally

Manon Verbockhaven,Guillaume Charpiat

Machine learning tasks are generally formulated as optimization problems, where one searches for an optimal function within a certain functional space. In practice, parameterized functional spaces are considered, in order to be able to perform gradient descent. Typically, a neural network architecture is chosen and fixed, and its parameters (connection weights) are optimized, yielding an architecture-dependent result. This way of proceeding however forces the evolution of the function during training to lie within the realm of what is expressible with the chosen architecture, and prevents any optimization across possible architectures. Costly architectural hyper-parameter optimization is often performed to compensate for this. Instead, we propose to adapt the architecture on the fly during training. We show that the information about desirable architectural changes, due to expressivity bottlenecks when attempting to follow the functional gradient, can be extracted from the back-propagation. To do this, we propose a new mathematically well-grounded method to detect expressivity bottlenecks on the fly and solve them by adding suitable neurons when and where needed. Thus, while the standard approach requires large networks, in terms of number of neurons per layer, for expressivity and optimization reasons, we are able to start with very small neural networks and let them grow appropriately. As a proof of concept, we show convincing results on the MNIST dataset, matching large neural network accuracy, with competitive training time, while removing the need for standard architectural hyper-parameter optimization.

****

Scalable and Privacy-enhanced Graph Generative Model for Graph Neural Networks

Minji Yoon,Yue Wu,John Palowitch,Bryan Perozzi,Russ Salakhutdinov

As the field of Graph Neural Networks (GNN) continues to grow, it experiences a corresponding increase in the need for large, real-world datasets to train and test new GNN models on challenging, realistic problems. Unfortunately, such graph datasets are often generated from online, highly privacy-restricted ecosystems,

which makes research and development on these datasets hard, if not impossible. This greatly reduces the amount of benchmark graphs available to researchers, causing the field to rely only on a handful of publicly-available datasets. To address this dilemma, we introduce a novel graph generative model, Computation Graph Transformer (CGT) that can learn and reproduce the distribution of real-world graphs in a privacy-enhanced way. Our proposed model (1) generates effective benchmark graphs on which GNNs show similar task performance as on the source graphs, (2) scales to process large-scale real-world graphs, (3) guarantees privacy for end-users. Extensive experiments across a vast body of graph generative models show that only our model can successfully generate privacy-controlled, synthetic substitutes of large-scale real-world graphs that can be effectively used to evaluate GNN models.
****

Model ensemble instead of prompt fusion: a sample-specific knowledge transfer method for few-shot prompt tuning
XIANGYU PENG,Chen Xing,Prafulla Kumar Choubey,Chien-Sheng Wu,Caiming Xiong
Prompt tuning approaches, which learn task-specific soft prompts for a downstream task conditioning on frozen pre-trained models, have attracted growing interest due to its parameter efficiency. With large language models and sufficient training data, prompt tuning performs comparably to full-model tuning. However, with limited training samples in few-shot settings, prompt tuning fails to match the performance of full-model fine-tuning. In this work, we focus on improving the few-shot performance of prompt tuning by transferring knowledge from soft prompts of source tasks with abundant training samples. Recognizing the good generalization capabilities of ensemble methods in low-data regime, we first experiment and show that a simple ensemble of model predictions based on different source prompts, outperforms existing multi-prompt knowledge transfer approaches such as source prompt fusion in the few-shot setting. Motivated by this observation, we further investigate model ensembles and propose Sample-specific Ensemble of Source Models (SESoM). SESoM learns to adjust the contribution of each source model for each target sample separately when ensembling source model outputs. Through this way, SESoM inherits the superior generalization of ensemble methods and simultaneously captures the sample-specific competence of each source prompt. We conduct experiments across a diverse set of eight NLP tasks using models of different scales (T5-\{base, large, XL\}) and find that SESoM consistently outperforms the existing models of the same as well as larger parametric scale by a large margin.
****

Entropy-Regularized Model-Based Offline Reinforcement Learning
Soroush Ghandi,Maryam Tavakol
Model-based approaches to offline Reinforcement Learning (RL) aim to remedy the problem of sample complexity in offline learning via first estimating a pessimistic Markov Decision Process (MDP) from offline data, followed by freely exploring in the learned model for policy optimization. Recent advances in model-based RL techniques mainly rely on an ensemble of models to quantify the uncertainty of the empirical MDP which is leveraged to penalize out-of-distribution state-action pairs during the policy learning. However, the performance of ensembles for uncertainty quantification highly depends on how they are implemented in practice, which can be a limiting factor. In this paper, we propose a systematic way to measure the epistemic uncertainty and present \abbrv, an Entropy-regularized Model-based Offline RL approach, to provide a smooth error estimation when leaving the support of data toward uncertain areas. Subsequently, we optimize a single neural architecture that maximizes the likelihood of offline data distribution while regularizing the transitions that are outside of the data support. Empirical results demonstrate that our framework achieves competitive performance compared to state-of-the-art offline RL methods on D4RL benchmark datasets.
****

Sign and Basis Invariant Networks for Spectral Graph Representation Learning
Derek Lim,Joshua David Robinson,Lingxiao Zhao,Tess Smidt,Suvrit Sra,Haggai Maron,Stefanie Jegelka

We introduce SignNet and BasisNet---new neural architectures that are invariant to two key symmetries displayed by eigenvectors: (i) sign flips, since if v is an eigenvector then so is -v; and (ii) more general basis symmetries, which occur in higher dimensional eigenspaces with infinitely many choices of basis eigenvectors. We prove that under certain conditions our networks are universal, i.e., they can approximate any continuous function of eigenvectors with the desired invariances. When used with Laplacian eigenvectors, our networks are provably more expressive than existing spectral methods on graphs; for instance, they subsume all spectral graph convolutions, certain spectral graph invariants, and previously proposed graph positional encodings as special cases. Experiments show that our networks significantly outperform existing baselines on molecular graph regression, learning expressive graph representations, and learning neural fields on triangle meshes. Our code is available at https://github.com/cptq/SignNet-BasisNet.

****

Diffusing Graph Attention
Daniel Glickman,Eran Yahav
The dominant paradigm for machine learning on graphs uses Message Passing Graph Neural Networks~(MP-GNNs), in which node representations are updated by aggregating information in their local neighborhood. Recently, there have been increasingly more attempts to adapt the Transformer architecture to graphs in an effort to solve some known limitations of MP-GNN. A challenging aspect of designing Graph Transformers is integrating the arbitrary graph structure into the architecture. We propose \emph{Graph Diffuser}~(GD) to address this challenge. GD learns to extract structural and positional relationships between distant nodes in the graph, which it then uses to direct the Transformer's attention and node representation. We demonstrate that existing GNNs and Graph Transformers struggle to capture long-range interactions and how Graph Diffuser does so while admitting intuitive visualizations. Experiments on eight benchmarks show Graph Diffuser to be a highly competitive model, outperforming the state-of-the-art in a diverse set of domains.

****

Sequential Latent Variable Models for Few-Shot High-Dimensional Time-Series Forecasting
Xiajun Jiang,Ryan Missel,Zhiyuan Li,Linwei Wang
Modern applications increasingly require learning and forecasting latent dynamics from high-dimensional time-series. Compared to univariate time-series forecasting, this adds a new challenge of reasoning about the latent dynamics of an unobserved abstract state. Sequential latent variable models (LVMs) present an attractive solution, although existing works either struggle with long-term forecasting or have difficulty learning across diverse dynamics. In this paper, we first present a conceptual framework of sequential LVMs to unify existing works, contrast their fundamental limitations, and identify an intuitive solution to long-term forecasting for diverse dynamics via meta-learning. We then present the first framework of few-shot forecasting for high-dimensional time-series: instead of learning a single dynamic function, we leverage data of diverse dynamics and learn to adapt latent dynamic functions to few-shot support series. This is realized via Bayesian meta-learning underpinned by: 1) a latent dynamic function conditioned on knowledge derived from few-shot support series, and 2) a meta-model that learns to extract such dynamic-specific knowledge via feed-forward embedding of support set. We compared the presented framework with a comprehensive set of baseline models trained 1) globally on the large meta-training set with diverse dynamics, and 2) individually on single dynamics, both with and without fine-tuning to k-shot support series used by the meta-models. We demonstrated that the presented framework is agnostic to the latent dynamic function of choice and, at meta-test time, is able to forecast for new dynamics given variable-shot of support series.

****

Code Translation with Compiler Representations
Marc Szafraniec,Baptiste Roziere,Hugh James Leather,Patrick Labatut,Francois Cha

rton,Gabriel Synnaeve

In this paper, we leverage low-level compiler intermediate representations (IR) code translation. Traditional transpilers rely on syntactic information and handcrafted rules, which limits their applicability and produces unnatural-looking code. Applying neural machine translation (NMT) approaches to code has successfully broadened the set of programs on which one can get a natural-looking translation. However, they treat the code as sequences of text tokens, and still do not differentiate well enough between similar pieces of code which have different semantics in different languages. The consequence is low quality translation, reducing the practicality of NMT, and stressing the need for an approach significantly increasing its accuracy. Here we propose to augment code translation with IRs, specifically LLVM IR, with results on the C++, Java, Rust, and Go languages. Our method improves upon the state of the art for unsupervised code translation, increasing the number of correct translations by 11% on average, and up to 79% for the Java → Rust pair with greedy decoding. With beam search, it increases the number of correct translations by 5.5% in average. We extend previous test sets for code translation, by adding hundreds of Go and Rust functions. Additionally, we train models with high performance on the problem of IR decompilation, generating programming source code from IR, and study using IRs as intermediary pivot for translation.

****

GAIN: On the Generalization of Instructional Action Understanding

Junlong Li,Guangyi Chen,Yansong Tang,Jinan Bao,Kun Zhang,Jie Zhou,Jiwen Lu

Despite the great success achieved in instructional action understanding by deep learning and mountainous data, deploying trained models to the unseen environment still remains a great challenge, since it requires strong generalizability of models from in-distribution training data to out-of-distribution (OOD) data. In this paper, we introduce a benchmark, named GAIN, to analyze the GeneralizAbility of INstructional action understanding models. In GAIN, we reassemble steps of existing instructional video training datasets to construct the OOD tasks and then collect the corresponding videos. We evaluate the generalizability of models trained on in-distribution datasets with the performance on OOD videos and observe a significant performance drop. We further propose a simple yet effective approach, which cuts off the excessive contextual dependency of action steps by performing causal inference, to provide a potential direction for enhancing the OOD generalizability. In the experiments, we show that this simple approach can improve several baselines on both instructional action segmentation and detection tasks. We expect the introduction of the GAIN dataset will promote future in-depth research on the generalization of instructional video understanding.

****

Deep Reinforcement learning on Adaptive Pairwise Critic and Asymptotic Actor

Huihui Zhang

Maximum entropy deep reinforcement learning has displayed great potential on a range of challenging continuous tasks. The maximum entropy is able to encourage policy exploration, however, it has a tradeoff between the efficiency and stability, especially when employed on large-scale tasks with high state and action dimensionality.

Sometimes the temperature hyperparameter of maximum entropy term is limited to remain stable at the cost of slower and lower convergence.

Besides, the function approximation errors existing in actor-critic learning are known to induce estimation errors and suboptimal policies.

In this paper, we propose an algorithm based on adaptive pairwise critics, and adaptive asymptotic maximum entropy combined.

Specifically, we add a trainable state-dependent weight factor to build an adaptive pairwise target Q-value to serve as the surrogate policy objective.

Then we adopt a state-dependent adaptive temperature to smooth the entropy policy exploration, which introduces an asymptotic maximum entropy.

The adaptive pairwise critics can effectively improve the value estimation, preventing overestimation or underestimation errors. Meanwhile, the adaptive asymptotic entropy can adapt to the tradeoff between efficiency and stability, which pr ovides more exploration and flexibility.

We evaluate our method on a set of Gym tasks, and the results show that the proposed algorithms have better performance than several baselines on continuous control.

****

Model-based Value Exploration in Actor-critic Deep Reinforcement Learning

Huihui Zhang

Off-policy method has demonstrated great potential on model-free deep reinforcement learning due to the sample-efficient advantage. However, it suffers extra instability due to some mismatched distributions from observations. Model-free on-policy counterparts usually have poor sample efficiency. Model-based algorithms, in contrast, are highly dependent on the goodness of expert demonstrations or learned dynamics.

In this work, we propose a method which involves training the dynamics to accelerate and gradually stabilize learning without adding sample-complexity. The dynamics model prediction can provide effective target value exploration, which is essentially different from the methods on-policy exploration, by adding valid diversity of transitions.

Despite the existence of model bias, the model-based prediction can avoid the overestimation and distribution mismatch errors in off-policy learning, as the learned dynamics model is asymptotically accurate.

Besides, to generalize the solution to large-scale reinforcement learning problems, we use global gaussian and deterministic function approximation to model the transition probability and reward function, respectively. To minimize the negative impact of potential model bias brought by the estimated dynamics, we adopt one-step global prediction for the model-based part of target value. By analyses and proofs, we show how the model-based prediction provides value exploration and asymptotical performance to the overall network. It can also be concluded that the convergence of proposed algorithm only depends on the accuracy of learnt dynamics model.

****

Omnigrok: Grokking Beyond Algorithmic Data

Ziming Liu,Eric J Michaud,Max Tegmark

Grokking, the unusual phenomenon for algorithmic datasets where generalization happens long after overfitting the training data, has remained elusive. We aim to understand grokking by analyzing the loss landscapes of neural networks, identifying the mismatch between training and test losses as the cause for grokking. We refer to this as the "LU mechanism" because training and test losses (against model weight norm) typically resemble "L" and "U", respectively. This simple mechanism can nicely explain many aspects of grokking: data size dependence, weight decay dependence, the emergence of representations, etc. Guided by the intuitive picture, we are able to induce grokking on tasks involving images, language and molecules, although the grokking signals are sometimes less dramatic. We attribute the dramatic nature of grokking for algorithmic datasets to representation learning.

****

ManyDG: Many-domain Generalization for Healthcare Applications

Chaoqi Yang,M Brandon Westover,Jimeng Sun

The vast amount of health data has been continuously collected for each patient, providing opportunities to support diverse healthcare predictive tasks such as seizure detection and hospitalization prediction. Existing models are mostly trained on other patients’ data and evaluated on new patients. Many of them might suffer from poor generalizability. One key reason can be overfitting due to the unique information related to patient identities and their data collection environments, referred to as patient covariates in the paper. These patient covariates usually do not contribute to predicting the targets but are often difficult to remove. As a result, they can bias the model training process and impede generalization. In healthcare applications, most existing domain generalization methods assume a small number of domains. In this paper, considering the diversity of patient covariates, we propose a new setting by treating each patient as a separa te domain (leading to many domains). We develop a new domain generalization method ManyDG, that can scale to such many-domain problems. Our method identifies the patient do- main covariates by mutual reconstruction, and removes them via an orthogonal projection step. Extensive experiments show that ManyDG can boost the generalization performance on multiple real-world healthcare tasks (e.g., 3.7% Jaccard improvements on MIMIC drug recommendation) and support realistic but challenging settings such as insufficient data and continuous learning. The code is available at https://github.com/ycq091044/ManyDG.

****

Adversarial Detector for Decision Tree Ensembles Using Representation Learning

Gal Braun,Lior Rokach

Research on adversarial evasion attacks focuses mainly on neural network models. Among other reasons, this is because of their popularity in certain fields (e.g., computer vision and NLP) and the models' properties, making it easier to search for adversarial examples with minimal input change. Decision trees and tree ensembles are still very popular due to their high
performance in fields dominated by tabular data and their explainability. In recent years, several works have defined new adversarial attacks targeting decision trees and tree ensembles. As a result, several papers were published focusing on robust versions of tree ensembles. This research aims to create an adversarial detector for attacks on an ensemble of decision trees. While several previous works have demonstrated the generation of more robust tree ensembles, the process of considering evasion attacks during ensemble generation can affect model performance. We demonstrate a method to detect adversarial samples without affecting either the target model structure or its original performance. We showed that by using representation learning based on the structure of the trees, we achieved better detection rates than the state-of-the-art technique and better than using the original representation of the dataset to train an adversarial detector.

****

Learning with Instance-Dependent Label Noise: Balancing Accuracy and Fairness

Donna Tjandra,Jenna Wiens

Incorrect labels hurt model performance when the model overfits to noise. Many state-of-the-art approaches that address label noise assume that label noise is independent from the input features. In practice, however, label noise is often feature or instance-\textit{dependent}, and therefore is biased (i.e., some instances are more likely to be mislabeled than others). Approaches that ignore this dependence can produce models with poor discriminative performance, and depending on the task, can exacerbate issues around fairness. In light of these limitations, we propose a two-stage approach to learn from datasets with instance-dependent label noise. Our approach utilizes \textit{anchor points}, a small subset of data for which we know the ground truth labels. On many tasks, our approach leads to consistent improvements over the state-of-the-art in discriminative performance (AUROC) while balancing model fairness (area under the equalized odds curve, AUEOC). For example, when predicting acute respiratory failure onset on the MIMIC-III dataset, the harmonic mean of the AUROC and AUEOC of our approach is 0.84 (SD 0.01) while that of the next best baseline is 0.81 (SD 0.01). Overall, our approach leads to learning more accurate and fair models compared to existing approaches in the presence of instance-dependent label noise.

****

Flow Annealed Importance Sampling Bootstrap

Laurence Illing Midgley,Vincent Stimper,Gregor N. C. Simm,Bernhard Schölkopf,José Miguel Hernández-Lobato

Normalizing flows are tractable density models that can approximate complicated target distributions, e.g. Boltzmann distributions of physical systems. However, current methods for training flows either suffer from mode-seeking behavior, use samples from the target generated beforehand by expensive MCMC methods, or use stochastic losses that have high variance. To avoid these problems, we augment flows with annealed importance sampling (AIS) and minimize the mass-covering α-divergence with $\alpha=2$, which minimizes importance weight variance. Our method, Flow AIS Bootstrap (FAB), uses AIS to generate samples in regions wher e the flow is a poor approximation of the target, facilitating the discovery of new modes. We apply FAB to multimodal targets and show that we can approximate them very accurately where previous methods fail. To the best of our knowledge, we are the first to learn the Boltzmann distribution of the alanine dipeptide molecule using only the unnormalized target density, without access to samples generated via Molecular Dynamics (MD) simulations: FAB produces better results than training via maximum likelihood on MD samples while using 100 times fewer target evaluations. After reweighting the samples, we obtain unbiased histograms of dihedral angles that are almost identical to the ground truth.

****

DecAF: Joint Decoding of Answers and Logical Forms for Question Answering over Knowledge Bases

Donghan Yu,Sheng Zhang,Patrick Ng,Henghui Zhu,Alexander Hanbo Li,Jun Wang,Yiqun Hu,William Yang Wang,Zhiguo Wang,Bing Xiang

Question answering over knowledge bases (KBs) aims to answer natural language questions with factual information such as entities and relations in KBs. Previous methods either generate logical forms that can be executed over KBs to obtain final answers or predict answers directly. Empirical results show that the former often produces more accurate answers, but it suffers from non-execution issues due to potential syntactic and semantic errors in the generated logical forms. In this work, we propose a novel framework DecAF that jointly generates both logical forms and direct answers, and then combines the merits of them to get the final answers. Moreover, different from most of the previous methods, DecAF is based on simple free-text retrieval without relying on any entity linking tools --- this simplification eases its adaptation to different datasets. DecAF achieves new state-of-the-art accuracy on WebQSP, FreebaseQA, and GrailQA benchmarks, while getting competitive results on the ComplexWebQuestions benchmark.

****

NANSY++: Unified Voice Synthesis with Neural Analysis and Synthesis

Hyeong-Seok Choi,Jinhyeok Yang,Juheon Lee,Hyeongju Kim

Various applications of voice synthesis have been developed independently despite the fact that they generate “voice” as output in common. In addition, most of the voice synthesis models still require a large number of audio data paired with annotated labels (e.g., text transcription and music score) for training. To this end, we propose a unified framework of synthesizing and manipulating voice signals from analysis features, dubbed NANSY++. The backbone network of NANSY++ is trained in a self-supervised manner that does not require any annotations paired with audio. After training the backbone network, we efficiently tackle four voice applications - i.e. voice conversion, text-to-speech, singing voice synthesis, and voice designing - by partially modeling the analysis features required for each task. Extensive experiments show that the proposed framework offers competitive advantages such as controllability, data efficiency, and fast training convergence, while providing high quality synthesis. Audio samples: tinyurl.com/8tnsy3uc.

****

Causality Compensated Attention for Contextual Biased Visual Recognition

Ruyang Liu,Jingjia Huang,Thomas H. Li,Ge Li

Visual attention does not always capture the essential object representation desired for robust predictions. Attention modules tend to underline not only the target object but also the common co-occurring context that the module thinks helpful in the training. The problem is rooted in the confounding effect of the context leading to incorrect causalities between objects and predictions, which is further exacerbated by visual attention. In this paper, to learn causal object features robust for contextual bias, we propose a novel attention module named Interventional Dual Attention (IDA) for visual recognition. Specifically, IDA adopts two attention layers with multiple sampling intervention, which compensates the attention against the confounder context. Note that our method is model-agnostic and thus can be implemented on various backbones. Extensive experiments show our model obtains significant improvements in classification and detection with lower computation. In particular, we achieve the state-of-the-art results in mul ti-label classification on MS-COCO and PASCAL-VOC.

****

A unified optimization framework of ANN-SNN Conversion: towards optimal mapping from activation values to firing rates

Haiyan Jiang,Srinivas Anumasa,Giulia De Masi,Huan Xiong,Bin Gu

Spiking Neural Networks (SNNs) have attracted great attention as a primary candidate for running large-scale deep artificial neural networks (ANNs) in real-time due to their distinctive properties of energy-efficient and event-driven fast computation. Training an SNN directly from scratch is usually difficult because of the discreteness of spikes. Converting an ANN to an SNN, i.e., ANN-SNN conversion, is an alternative method to obtain deep SNNs.

The performance of the converted SNN is determined by both the ANN performance and the conversion error. The existing ANN-SNN conversion methods usually redesign the ANN with a new activation function instead of the regular ReLU, train the tailored ANN and convert it to an SNN. The performance loss between the regular ANN with ReLU and the tailored ANN has never been considered, which will be inherited to the converted SNN.

In this work, we formulate the ANN-SNN conversion as a unified optimization problem which considers the performance loss between the regular ANN and the tailored ANN, as well as the conversion error simultaneously. Following the unified optimization framework, we propose the SlipReLU activation function to replace the regular ReLU activation function in the tailored ANN. The SlipReLU is a weighted sum of the threhold-ReLU and the step function, which improves the performance of either as an activation function alone.

The SlipReLU method covers a family of activation functions mapping from activation values in source ANNs to firing rates in target SNNs; most of the state-of-the-art optimal ANN-SNN conversion methods are special cases of our proposed SlipReLU method. We demonstrate through two theorems that the expected conversion error between SNNs and ANNs can theoretically be zero on a range of shift values $\delta \in [-\frac{1}{2},\frac{1}{2}]$ rather than a fixed shift term $\frac{1}{2}$, enabling us to achieve converted SNNs with high accuracy and ultra-low latency. We evaluate our proposed SlipReLU method on CIFAR-10 dataset, and the results show that the SlipReLU outperforms the state-of-the-art ANN-SNN conversion in both accuracy and latency. To our knowledge, this is the first work to explore high-performance ANN-SNN conversion method considering the ANN performance and the conversion error simultaneously.

****

Multi-Objective Reinforcement Learning: Convexity, Stationarity and Pareto Optimality

Haoye Lu,Daniel Herman,Yaoliang Yu

In recent years, single-objective reinforcement learning (SORL) algorithms have received a significant amount of attention and seen some strong results. However, it is generally recognized that many practical problems have intrinsic multi-objective properties that cannot be easily handled by SORL algorithms. Although there have been many multi-objective reinforcement learning (MORL) algorithms proposed, there has been little recent exploration of the fundamental properties of the spaces we are learning in. In this paper, we perform a rigorous analysis of policy induced value functions and use the insights to distinguish three views of Pareto optimality. The results imply the convexity of the induced value function's range for stationary policies and suggest that any point of its Pareto front can be achieved by training a policy using linear scalarization (LS). We show the problem that leads to the suboptimal performance of LS can be solved by adding strongly concave terms to the immediate rewards, which motivates us to propose a new vector reward-based Q-learning algorithm, CAPQL. Combined with an actor-critic formulation, our algorithm achieves state-of-the-art performance on multiple MuJoCo tasks in the preference agnostic setting. Furthermore, we empirically show that, in contrast to other LS-based algorithms, our approach is significantly more stable, achieving similar results across various random seeds.

****

Continual Unsupervised Disentangling of Self-Organizing Representations

Zhiyuan Li,Xiajun Jiang,Ryan Missel,Prashnna Kumar Gyawali,Nilesh Kumar,Linwei Wang

Limited progress has been made in continual unsupervised learning of representations, especially in reusing, expanding, and continually disentangling learned semantic factors across data environments. We argue that this is because existing approaches treat continually-arrived data independently, without considering how they are related based on the underlying semantic factors. We address this by a new generative model describing a topologically-connected mixture of spike-and-slab distributions in the latent space, learned end-to-end in a continual fashion via principled variational inference. The learned mixture is able to automatically discover the active semantic factors underlying each data environment and to accumulate their relational structure based on that. This distilled knowledge of different data environments can further be used for generative replay and guiding continual disentangling of new semantic factors. We tested the presented method on a split version of 3DShapes to provide the first quantitative disentanglement evaluation of continually learned representations, and further demonstrated its ability to continually disentangle new representations in benchmark datasets.

****

Inducing Gaussian Process Networks

Alessandro Tibo,Thomas Dyhre Nielsen

Gaussian processes (GPs) are powerful but computationally expensive machine learning models, requiring an estimate of the kernel covariance matrix for every prediction. In large and complex domains, such as graphs, sets, or images, the choice of suitable kernel can also be non-trivial to determine, providing an additional obstacle to the learning task. Over the last decade, these challenges have resulted in significant advances being made in terms of scalability and expressivity, exemplified by, e.g., the use of inducing points and neural network kernel approximations.

In this paper, we propose inducing Gaussian process networks (IGN), a simple framework for simultaneously learning the feature space as well as the inducing points. The inducing points, in particular, are learned directly in the feature space, enabling a seamless representation of complex structured domains while also facilitating scalable gradient-based learning methods.

We consider both regression and (binary) classification tasks and report on experimental results for real-world data sets showing that IGNs provide significant advances over state-of-the-art methods. We also demonstrate how IGNs can be used to effectively model complex domains using neural network architectures.

****

Monotonicity and Double Descent in Uncertainty Estimation with Gaussian Processes

Liam Hodgkinson,Chris van der Heide,Fred Roosta,Michael W. Mahoney

The quality of many modern machine learning models improves as model complexity increases, an effect that has been quantified—for predictive performance—with the non-monotonic double descent learning curve. Here, we address the overarching question: is there an analogous theory of double descent for models which estimate uncertainty? We provide a partially affirmative and partially negative answer in the setting of Gaussian processes (GP). Under standard assumptions, we prove that higher model quality for optimally-tuned GPs (including uncertainty prediction) under marginal likelihood is realized for larger input dimensions, and therefore exhibits a monotone learning curve. After showing that marginal likelihood does not naturally exhibit double descent in the input dimension, we highlight related forms of posterior predictive loss that do. Finally, we verify empirically that our results hold for real data, beyond our considered assumptions, and explore unusual consequences involving synthetic covariates.

****

Fooling SHAP with Stealthily Biased Sampling

gabriel laberge,Ulrich Aïvodji,Satoshi Hara,Mario Marchand,Foutse Khomh

SHAP explanations aim at identifying which features contribute the most to the difference in model prediction at a specific input versus
a background distribution. Recent studies have shown that they can be manipulated by malicious adversaries to produce arbitrary desired
explanations. However, existing attacks focus solely on altering the black-box model itself. In this paper, we propose a complementary family
of attacks that leave the model intact and manipulate SHAP explanations using stealthily biased sampling of the data points used to approximate expectations w.r.t the background distribution. In the context of fairness audit, we show that our attack can reduce the importance of a sensitive feature when explaining the difference in outcomes between groups while remaining undetected. More precisely, experiments performed on real-world datasets showed that our attack could yield up to a 90\% relative decrease in amplitude of the sensitive feature attribution. These results highlight the manipulability of SHAP explanations and encourage auditors to treat them with skepticism.
****
Asynchronous Gradient Play in Zero-Sum Multi-agent Games
Ruicheng Ao,Shicong Cen,Yuejie Chi
Finding equilibria via gradient play in competitive multi-agent games has been attracting a growing amount of attention in recent years, with emphasis on designing efficient strategies where the agents operate in a decentralized and symmetric manner with guaranteed convergence. While significant efforts have been made in understanding zero-sum two-player matrix games, the performance in zero-sum multi-agent games remains inadequately explored, especially in the presence of delayed feedbacks, leaving the scalability and resiliency of gradient play open to questions. In this paper, we make progress by studying asynchronous gradient plays in zero-sum polymatrix games under delayed feedbacks. We first establish that the last iterate of entropy-regularized optimistic multiplicative weight updates (OMWU) method converges linearly to the quantal response equilibrium (QRE), the solution concept under bounded rationality, in the absence of delays. The linear convergence continues to hold even when the feedbacks are randomly delayed under mild statistical assumptions, albeit at a slower rate. Moving beyond random delays, we further demonstrate entropy-regularized OMWU with two-timescale learning rates enjoys faster last-iterate convergence under fixed delays, and continues to converge provably even when the delays are arbitrarily bounded. Our methods also lead to finite-time guarantees to approximate the Nash equilibrium (NE) by moderating the amount of regularization. To the best of our knowledge, this work is the first that aims to understand asynchronous gradient play in zero-sum polymatrix games under a wide range of delay assumptions.

****
Novel View Synthesis with Diffusion Models
Daniel Watson,William Chan,Ricardo Martin Brualla,Jonathan Ho,Andrea Tagliasacchi,Mohammad Norouzi
We present 3DiM (pronounced "three-dim"), a diffusion model for 3D novel view synthesis from as few as a single image. The core of 3DiM is an image-to-image diffusion model -- 3DiM takes a single reference view and their poses as inputs, and generates a novel view via diffusion. 3DiM can then generate a full 3D consistent scene following our novel stochastic conditioning sampler: the output frames of the scene are generated autoregressively, and during the reverse diffusion process of each individual frame, we select a random conditioning frame from the set of previous frames at each denoising step. We demonstrate that stochastic conditioning yields much more 3D consistent results compared to the naive sampling process which only conditions on a single previous frame. We compare 3DiMs to prior work on the SRN ShapeNet dataset, demonstrating that 3DiM's generated videos from a single view achieve much higher fidelity while being approximately 3D consistent. We also introduce a new evaluation methodology, 3D consistency scoring, to measure the 3D consistency of a generated object by training a neural field on the model's output views. 3DiMs are geometry free, do not rely on hyper-networks or test-time optimization for novel view synthesis, and allow a single mod

el to easily scale to a large number of scenes.
****

DM-NeRF: 3D Scene Geometry Decomposition and Manipulation from 2D Images
Bing WANG,Lu Chen,Bo Yang

In this paper, we study the problem of 3D scene geometry decomposition and manipulation from 2D views. By leveraging the recent implicit neural representation techniques, particularly the appealing neural radiance fields, we introduce an object field component to learn unique codes for all individual objects in 3D space only from 2D supervision. The key to this component is a series of carefully designed loss functions to enable every 3D point, especially in non-occupied space, to be effectively optimized even without 3D labels. In addition, we introduce an inverse query algorithm to freely manipulate any specified 3D object shape in the learned scene representation. Notably, our manipulation algorithm can explicitly tackle key issues such as object collisions and visual occlusions. Our method, called DM-NeRF, is among the first to simultaneously reconstruct, decompose, manipulate and render complex 3D scenes in a single pipeline. Extensive experiments on three datasets clearly show that our method can accurately decompose all 3D objects from 2D views, allowing any interested object to be freely manipulated in 3D space such as translation, rotation, size adjustment, and deformation.
****

Robust Neural ODEs via Contractivity-promoting Regularization
Liang Xu,Muhammad Zakwan,Giancarlo Ferrari-Trecate

Neural networks can be fragile to input noise and adversarial attacks. In this work, we consider Neural Ordinary Differential Equations (NODEs) – a family of continuous-depth neural networks represented by dynamical systems - and propose to use contraction theory to improve their robustness. A dynamical system is contractive if two trajectories starting from different initial conditions converge to each other exponentially fast. Contractive NODEs can enjoy increased robustness as slight perturbations of the features do not cause a significant change in the output. Contractivity can be induced during training by using a regularization term involving the Jacobian of the system dynamics. To reduce the computational burden, we show that it can also be promoted using carefully selected weight regularization terms for a class of NODEs with slope-restricted activation functions, including convolutional networks commonly used in image classification. The performance of the proposed regularizers is illustrated through benchmark image classification tasks on MNIST and FashionMNIST datasets, where images are corrupted by different kinds of noise and attacks.
****

Analyzing the Effects of Classifier Lipschitzness on Explainers
Zulqarnain Khan,Aria Masoomi,Davin Hill,Jennifer Dy

Machine learning methods are getting increasingly better at making predictions, but at the same time they are also becoming more complicated and less transparent. As a result, explainers are often relied on to provide interpretability to these \textit{black-box} prediction models. As crucial diagnostics tools, it is important that these explainers themselves are reliable. In this paper we focus on one particular aspect of reliability, namely that an explainer should give similar explanations for similar data inputs. We formalize this notion by introducing and defining \textit{explainer astuteness}, analogous to astuteness of classifiers. Our formalism is inspired by the concept of \textit{probabilistic Lipschitzness}, which captures the probability of local smoothness of a function. For a variety of explainers (e.g., SHAP, RISE, CXPlain), we provide lower bound guarantees on the astuteness of these explainers given the Lipschitzness of the prediction function. These theoretical results imply that locally smooth prediction functions lend themselves to locally robust explanations. We evaluate these results empirically on simulated as well as real datasets.
****

Complex-Target-Guided Open-Domain Conversation based on offline reinforcement learning
Haibo Huang,Ying Tan,Haoxing Zhang,Lei Yu

Previous target-guided open-domain dialogue systems mostly take one keyword as the target, which has great limitations and cannot characterize the dialogue target well. In this paper, we introduce a new target representation model which uses a verb-noun pair to represent a complex-target. To this end, we implement a new dialogue guide procedure with Verb graph and Noun graph construction, dialogue encoder, verb-noun choose model and response generator. Machine metrics and human evaluation both show that our model outperforms previous target-guided dialogue system. In addition, different from previous target-guided dialogue systems which use online reinforcement learning to make decisions, we integrate an offline reinforcement learning method to gradually reduce the training time with a high performance.

****

Trading Information between Latents in Hierarchical Variational Autoencoders
Tim Z. Xiao,Robert Bamler

Variational Autoencoders (VAEs) were originally motivated as probabilistic generative models in which one performs approximate Bayesian inference. The proposal of β-VAEs breaks this interpretation and generalizes VAEs to application domains beyond generative modeling (e.g., representation learning, clustering, or lossy data compression) by introducing an objective function that allows practitioners to trade off between the information content ("bit rate") of the latent representation and the distortion of reconstructed data. In this paper, we reconsider this rate/distortion trade-off in the context of hierarchical VAEs, i.e., VAEs with more than one layer of latent variables. We propose a method to control each layer's contribution to the rate independently. We identify the most general class of inference models to which our proposed method is applicable, and we derive theoretical bounds on the performance of downstream tasks as functions of the individual layers' rates. Our experiments demonstrate that the proposed method allows us to better tune hierarchical VAEs for a diverse set of practical use cases.

****

VC Theoretical Explanation of Double Descent
Eng Hock Lee,Vladimir Cherkassky

There has been growing interest in generalization performance of large multilayer neural networks that can be trained to achieve zero training error, while generalizing well on test data. This regime is known as ‘second descent’ and it appears to contradict the conventional view that optimal model complexity should reflect an optimal balance between underfitting and overfitting, i.e., the bias-variance tradeoff.

This paper presents a VC-theoretical analysis of double descent and shows that it can be fully explained by classical VC-generalization bounds. We illustrate an application of analytic VC-bounds for modeling double descent for classification problems, using empirical results for several learning methods, such as SVM, Least Squares, and Multilayer Perceptron classifiers. In addition, we discuss several reasons for the misinterpretation of VC-theoretical results in Deep Learning community.

****

Enhance Local Consistency for Free: A Multi-Step Inertial Momentum Approach
Yixing Liu,Yan Sun,Li Shen,Baoyuan Wu,Zhengtao Ding,Dacheng Tao

Federated learning (FL), as a collaborative distributed training paradigm with several edge computing devices under the coordination of a centralized server, is plagued by inconsistent local stationary points due to the heterogeneity of the local partial participation clients, which precipitates the local client-drifts problems and sparks off the unstable and slow convergence, especially on the aggravated heterogeneous dataset. To address these issues, we propose a novel federated learning algorithm, named FedMIM, which adopts the multi-step inertial momentum on the edge devices and enhances the local consistency for free during the training to improve the robustness of the heterogeneity. Specifically, we incorporate the weighted global gradient estimations as the inertial correction terms to guide both the local iterates and stochastic gradient estimation, which can reckon the global objective optimization on the edges' heterogeneous dataset nat urally and maintain the demanding consistent iteration locally. Theoretically, we show that FedMIM achieves the $\mathcal{O}\big({1}/{\sqrt{SKT}}\big)$ convergence rate with a linear speedup property with respect to the number of selected clients S and proper local interval K in each communication round under the nonconvex setting. Empirically, we conduct comprehensive experiments on various real-world datasets and demonstrate the efficacy of the proposed FedMIM against several state-of-the-art baselines.

****

SYNG4ME: Model Evaluation using Synthetic Test Data

Boris van Breugel,Nabeel Seedat,Fergus Imrie,Mihaela van der Schaar

Model evaluation is a crucial step in ensuring reliable machine learning systems. Currently, predictive models are evaluated on held-out test data, quantifying aggregate model performance. Limitations of available test data make it challenging to evaluate model performance on small subgroups or when the environment changes. Synthetic test data provides a unique opportunity to address this challenge; instead of evaluating predictive models on real data, we propose to use synthetic data. This brings two advantages. First, supplementing and increasing the amount of evaluation data can lower the variance of model performance estimates compared to evaluation on the original test data. This is especially true for local performance evaluation in low-density regions, e.g. minority or intersectional groups. Second, generative models can be conditioned as to induce a shift in the synthetic data distribution, allowing us to evaluate how supervised models could perform in different target settings. In this work, we propose SYNG4ME: an automated suite of synthetic data generators for model evaluation. By generating smart synthetic data sets, data practitioners have a new tool for exploring how supervised models may perform on subgroups of the data, and how robust methods are to distributional shifts. We show experimentally that SYNG4ME achieves more accurate performance estimates compared to using the test data alone.

****

Take One Gram of Neural Features, Get Enhanced Group Robustness

Simon Roburin,Charles Corbière,Gilles Puy,Nicolas THOME,Mathieu Aubry,Renaud Marlet,Patrick Perez

Predictive performance of machine learning models trained with empirical risk minimization (ERM) can degrade considerably under distribution shifts. In particular, the presence of spurious correlations in training datasets leads ERM-trained models to display high loss when evaluated on minority groups not presenting such correlations in test sets. Extensive attempts have been made to develop methods improving worst-group robustness. However, they require group information for each training input or at least, a validation set with group labels to tune their hyperparameters, which may be expensive to get or unknown a priori. In this paper, we address the challenge of improving group robustness without group annotations during training. To this end, we propose to partition automatically the training dataset into groups based on Gram matrices of features extracted from an identification model and to apply robust optimization based on these pseudo-groups. In the realistic context where no group labels are available, our experiments show that our approach not only improves group robustness over ERM but also outperforms all recent baselines.

****

LMC: Fast Training of GNNs via Subgraph Sampling with Provable Convergence

Zhihao Shi,Xize Liang,Jie Wang

The message passing-based graph neural networks (GNNs) have achieved great success in many real-world applications.

However, training GNNs on large-scale graphs suffers from the well-known neighbor explosion problem, i.e., the exponentially increasing dependencies of nodes with the number of message passing layers. Subgraph-wise sampling methods---a promising class of mini-batch training techniques---discard messages outside the mini-batches in backward passes to avoid the neighbor explosion problem at the expense of gradient estimation accuracy. This poses significant challenges to their convergence analysis and convergence speeds, which seriously limits their reliable real-world applications. To address this challenge, we propose a novel subgra ph-wise sampling method with a convergence guarantee, namely Local Message Compensation (LMC). To the best of our knowledge, LMC is the {\it first} subgraph-wise sampling method with provable convergence. The key idea of LMC is to retrieve the discarded messages in backward passes based on a message passing formulation of backward passes. By efficient and effective compensations for the discarded messages in both forward and backward passes, LMC computes accurate mini-batch gradients and thus accelerates convergence. We further show that LMC converges to first-order stationary points of GNNs. Experiments on large-scale benchmark tasks demonstrate that LMC significantly outperforms state-of-the-art subgraph-wise sampling methods in terms of efficiency.

****

DEEPER-GXX: DEEPENING ARBITRARY GNNS
Lecheng Zheng,Dongqi Fu,Ross Maciejewski,Jingrui He

Recently, motivated by real applications, a major research direction in graph neural networks (GNNs) is to explore deeper structures.

For instance, the graph connectivity is not always consistent with the label distribution (e.g., the closest neighbors of some nodes are not from the same category). In this case, GNNs need to stack more layers, in order to find the same categorical neighbors in a longer path for capturing the class-discriminative information. However, two major problems hinder the deeper GNNs to obtain satisfactory performance, i.e., vanishing gradient and over-smoothing. On one hand, stacking layers makes the neural network hard to train as the gradients of the first few layers vanish. Moreover, when simply addressing vanishing gradient in GNNs, we discover the shading neighbors effect (i.e., stacking layers inappropriately distorts the non-IID information of graphs and degrade the performance of GNNs). On the other hand, deeper GNNs aggregate much more information from common neighbors such that individual node representations share more overlapping features, which makes the final output representations not discriminative (i.e., overly smoothed). In this paper, for the first time, we address both problems to enable deeper GNNs, and propose Deeper-GXX, which consists of the Weight-Decaying Graph Residual Connection module (WDG-ResNet) and Topology-Guided Graph Contrastive Loss (TGCL). Extensive experiments on real-world data sets demonstrate that Deeper-GXX outperforms state-of-the-art deeper baselines.

****

Music-to-Text Synaesthesia: Generating Descriptive Text from Music Recordings
Zhihuan Kuang,Shi Zong,Jianbing Zhang,Jiajun Chen,Hongfu Liu

In this paper, we consider a novel research problem, music-to-text synaesthesia. Different from the classical music tagging problem that classifies a music recording into pre-defined categories, the music-to-text synaesthesia aims to generate descriptive texts from music recordings for further understanding. Although this is a new and interesting application to the machine learning community, to our best knowledge, the existing music-related datasets do not contain the semantic description on music recordings and cannot serve the music-to-text synaesthesia task. In light of this, we collect a new dataset that contains 1,955 aligned pairs of classical music recordings and text descriptions. Based on this, we build a computational model to generate sentences that can describe the content of the music recording. To tackle the highly non-discriminative classical music, we design a group topology-preservation loss in our computational model, which considers more samples as a group reference and preserves the relative topology among different samples. Extensive experimental results qualitatively and quantitatively demonstrate the effectiveness of our proposed model over five heuristics or pre-trained competitive methods and their variants on our collected dataset.

****

ISAAC Newton: Input-based Approximate Curvature for Newton's Method
Felix Petersen,Tobias Sutter,Christian Borgelt,Dongsung Huh,Hilde Kuehne,Yuekai Sun,Oliver Deussen

We present ISAAC (Input-baSed ApproximAte Curvature), a novel method that conditions the gradient using selected second-order information and has an asymptotically vanishing computational overhead, assuming a batch size smaller than the number of neurons. We show that it is possible to compute a good conditioner based on only the input to a respective layer without a substantial computational overhead. The proposed method allows effective training even in small-batch stochastic regimes, which makes it competitive to first-order as well as second-order methods.

****

Learning Human-Compatible Representations for Case-Based Decision Support
Han Liu,Yizhou Tian,Chacha Chen,Shi Feng,Yuxin Chen,Chenhao Tan

Algorithmic case-based decision support provides examples to help human make sense of predicted labels and aid human in decision-making tasks. Despite the promising performance of supervised learning, representations learned by supervised models may not align well with human intuitions: what models consider as similar examples can be perceived as distinct by humans. As a result, they have limited effectiveness in case-based decision support. In this work, we incorporate ideas from metric learning with supervised learning to examine the importance of alignment for effective decision support. In addition to instance-level labels, we use human-provided triplet judgments to learn human-compatible decision-focused representations. Using both synthetic data and human subject experiments in multiple classification tasks, we demonstrate that such representation is better aligned with human perception than representation solely optimized for classification. Human-compatible representations identify nearest neighbors that are perceived as more similar by humans and allow humans to make more accurate predictions, leading to substantial improvements in human decision accuracies (17.8% in butterfly vs. moth classification and 13.2% in pneumonia classification).

****

Long-Tailed Learning Requires Feature Learning
Thomas Laurent,James von Brecht,Xavier Bresson

We propose a simple data model inspired from natural data such as text or images, and use it to study the importance of learning features in order to achieve good generalization. Our data model follows a long-tailed distribution in the sense that some rare and uncommon subcategories have few representatives in the training set. In this context we provide evidence that a learner succeeds if and only if it identifies the correct features, and moreover derive non-asymptotic generalization error bounds that precisely quantify the penalty that one must pay for not learning features.

****

Understanding Hindsight Goal Relabeling Requires Rethinking Divergence Minimization
Lunjun Zhang,Bradly C. Stadie

Hindsight goal relabeling has become a foundational technique for multi-goal reinforcement learning (RL). The idea is quite simple: any arbitrary trajectory can be seen as an expert demonstration for reaching the trajectory's end state. Intuitively, this procedure trains a goal-conditioned policy to imitate a sub-optimal expert. However, this connection between imitation and hindsight relabeling is not well understood. Modern imitation learning algorithms are described in the language of divergence minimization, and yet it remains an open problem how to recast hindsight goal relabeling into that framework. In this work, we develop a unified objective for goal-reaching that explains such a connection, from which we can derive goal-conditioned supervised learning (GCSL) and the reward function in hindsight experience replay (HER) from first principles. Experimentally, we find that despite recent advances in goal-conditioned behaviour cloning (BC), multi-goal Q-learning can still outperform BC-like methods; moreover, a vanilla combination of both actually hurts model performance. Under our framework, we study when BC is expected to help, and empirically validate our findings. Our work further bridges goal-reaching and generative modeling, illustrating the nuances and new pathways of extending the success of generative models to RL.

****

DoE2Vec: Representation Learning for Exploratory Landscape Analysis
Bas Van Stein,Fu Xing Long,Moritz A. Frenzel,Peter Krause,Markus Gitterle,Thomas Bäck

We propose DoE2Vec, a variational autoencoder (VAE)-based methodology to learn o ptimization landscape characteristics for downstream meta-learning tasks, e.g., automated selection of optimization algorithms. Principally, using large training data sets generated with a random function generator, DoE2Vec self-learns an informative latent representation for any design of experiments (DoE).

Unlike the classical exploratory landscape analysis (ELA) method, our approach does not require any feature engineering and is easily applicable for high dimensional search spaces. For validation, we inspect the quality of latent reconstructions and analyze the latent representations using different experiments.

The latent representations not only show promising potentials in identifying similar (cheap-to-evaluate) surrogate functions, but also can boost performances when being used complementary to the ELA features in classification tasks.

****

How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection?

Yifei Ming,Yiyou Sun,Ousmane Dia,Yixuan Li

Out-of-distribution (OOD) detection is a critical task for reliable machine learning. Recent advances in representation learning give rise to distance-based OOD detection, where testing samples are detected as OOD if they are relatively far away from the centroids or prototypes of in-distribution (ID) classes. However, prior methods directly take off-the-shelf contrastive losses that suffice for classifying ID samples, but are not optimally designed when test inputs contain OOD samples. In this work, we propose CIDER, a novel representation learning framework that exploits hyperspherical embeddings for OOD detection. CIDER jointly optimizes two losses to promote strong ID-OOD separability: a dispersion loss that promotes large angular distances among different class prototypes, and a compactness loss that encourages samples to be close to their class prototypes. We analyze and establish the unexplored relationship between OOD detection performance and the embedding properties in the hyperspherical space, and demonstrate the importance of dispersion and compactness. CIDER establishes superior performance, outperforming the latest rival by 19.36% in FPR95. Code is available at https://github.com/deeplearning-wisc/cider.

****

AnyDA: Anytime Domain Adaptation

Omprakash Chakraborty,Aadarsh Sahoo,Rameswar Panda,Abir Das

Unsupervised domain adaptation is an open and challenging problem in computer vision. While existing research shows encouraging results in addressing cross-domain distribution shift on common benchmarks, they are often constrained to testing under a specific target setting, limiting their impact for many real-world applications. In this paper, we introduce a simple yet effective framework for anytime domain adaptation that is executable with dynamic resource constraints to achieve accuracy-efficiency trade-offs under domain-shifts. We achieve this by training a single shared network using both labeled source and unlabeled data, with switchable depth, width and input resolutions on the fly to enable testing under a wide range of computation budgets. Starting with a teacher network trained from a label-rich source domain, we utilize bootstrapped recursive knowledge distillation as a nexus between source and target domains to jointly train the student network with switchable subnetworks. Experiments on multiple datasets well demonstrate the superiority of our approach over state-of-the-art methods.

****

Improving Deep Regression with Ordinal Entropy

Shihao Zhang,Linlin Yang,Michael Bi Mi,Xiaoxu Zheng,Angela Yao

In computer vision, it is often observed that formulating regression problems as a classification task yields better performance. We investigate this curious phenomenon and provide a derivation to show that classification, with the cross-entropy loss, outperforms regression with a mean squared error loss in its ability to learn high-entropy feature representations. Based on the analysis, we propose an ordinal entropy loss to encourage higher-entropy feature spaces while maintaining ordinal relationships to improve the performance of regression tasks. Experiments on synthetic and real-world regression tasks demonstrate the importance and benefits of increasing entropy for regression.

****

Revisiting Pretraining Objectives for Tabular Deep Learning
Ivan Rubachev,Artem Alekberov,Yury Gorishniy,Artem Babenko
Recent deep learning models for tabular data currently compete with the traditional ML models based on decision trees (GBDT). Unlike GBDT, deep models can additionally benefit from pretraining, which is a workhorse of DL for vision and NLP.

For tabular problems, several pretraining methods were proposed, but it is not entirely clear if pretraining provides consistent noticeable improvements and what method should be used, since the methods are often not compared to each other or comparison is limited to the simplest MLP architectures.

In this work, we aim to identify the best practices to pretrain tabular DL models that can be universally applied to different datasets and architectures.
Among our findings, we show that using the object target labels during the pretraining stage is beneficial for the downstream performance and advocate several target-aware pretraining objectives.
Overall, our experiments demonstrate that properly performed pretraining significantly increases the performance of tabular DL models, which often leads to their superiority over GBDTs.
****
OoD-Control: Out-of-Distribution Generalization for Adaptive UAV Flight Control
Yuxiao Duan,Jundong Zhou,Zhaoyu Zeng,Haoqi Zeng,Nanyang Ye
Data-driven control methods have demonstrated precise and agile control of Unmanned Aerial Vehicles (UAVs) over turbulence environments. However, they are relatively weak at taming the out-of-distribution (OoD) data, i.e., encountering the generalization problem when faced with unknown environments with different data distributions from the training set. Many studies have designed algorithms to reduce the impact of the OoD problem, a common but tricky problem in machine learning. To tackle the OoD generalization problem in control, we propose a theoretically guaranteed approach: OoD-Control. We provide proof that for any perturbation within some range on the states, the control error can be upper bounded by a constant. In this paper, we present our OoD-Control generalization algorithm for online adaptive flight control and execute it on two instances. Experiments show that systems trained by the proposed OoD-Control algorithm perform better in quite different environments from training. And the control method is extensible and pervasively applicable and can be applied to different dynamical models. OoD-Control is validated on UAV dynamic models, and we find it performs state-of-the-art in positioning stability and trajectory tracking problems.
****
Unified Discrete Diffusion for Simultaneous Vision-Language Generation
Minghui Hu,Chuanxia Zheng,Zuopeng Yang,Tat-Jen Cham,Heliang Zheng,Chaoyue Wang,Dacheng Tao,Ponnuthurai N. Suganthan
The recently developed discrete diffusion model performs extraordinarily well in generation tasks, especially in the text-to-image task, showing great potential for modeling multimodal signals. In this paper, we leverage these properties and present a unified multimodal generation model, which can perform text-based, image-based, and even vision-language simultaneous generation using a single model. Specifically, we unify the discrete diffusion process for multimodal signals by proposing a unified Markov transition matrix and a unified objective. Moreover, we design a multimodal mutual attention module to highlight the inter-modal linkages, which is vital for multimodal generation. Extensive experiments indicate that our proposed method can perform comparably to the state-of-the-art solutions in various generation tasks.
****
Take 5: Interpretable Image Classification with a Handful of Features
Thomas Norrenbrock,Marco Rudolph,Bodo Rosenhahn
Deep Neural Networks use thousands of mostly incomprehensible features to identify a single class, a decision no human can follow. We propose an interpretable sparse and low dimensional final decision layer in a deep neural network with measurable aspects of interpretability and demonstrate it on fine-grained image cla

ssification. We argue that a human can only understand the decision of a machine learning model, if the input features are interpretable and only very few of them are used for a single decision. For that matter, the final layer has to be sparse and – to make interpreting the features feasible – low dimensional. We call a model with a Sparse Low-Dimensional Decision “SLDD-Model”. We show that a SLDD-Model is easier to interpret locally and globally than a dense high-dimensional decision layer while being able to maintain competitive accuracy. Additionally, we propose a loss function that improves a model’s feature diversity and accuracy. Our interpretable SLDD-Model only uses 5 out of just 50 features per class, while maintaining 97% to 100% of the accuracy on four common benchmark datasets compared to the baseline model with 2048 features.

****

On the Fast Convergence of Unstable Reinforcement Learning Problems
Wang Zhang,Lam M. Nguyen,Subhro Das,Alexandre Megretski,Luca Daniel,Tsui-Wei Weng

For many of the reinforcement learning applications, the system is assumed to be inherently stable and with bounded reward, state and action space. These are key requirements for the optimization convergence of classical reinforcement learning reward function with discount factors. Unfortunately, these assumptions do not hold true for many real world problems such as an unstable linear–quadratic regulator (LQR). In this work, we propose new methods to stabilize and speed up the convergence of unstable reinforcement learning problems with the policy gradient methods. We provide theoretical insights on the efficiency of our methods. In practice, our method achieve good experimental results over multiple examples where the vanilla methods mostly fail to converge due to system instability.

****

Iterative Patch Selection for High-Resolution Image Recognition
Benjamin Bergner,Christoph Lippert,Aravindh Mahendran

High-resolution images are prevalent in various applications, such as autonomous driving and computer-aided diagnosis. However, training neural networks on such images is computationally challenging and easily leads to out-of-memory errors even on modern GPUs. We propose a simple method, Iterative Patch Selection (IPS), which decouples the memory usage from the input size and thus enables the processing of arbitrarily large images under tight hardware constraints. IPS achieves this by selecting only the most salient patches, which are then aggregated into a global representation for image recognition. For both patch selection and aggregation, a cross-attention based transformer is introduced, which exhibits a close connection to Multiple Instance Learning. Our method demonstrates strong performance and has wide applicability across different domains, training regimes and image sizes while using minimal accelerator memory. For example, we are able to finetune our model on whole-slide images consisting of up to 250k patches (>16 gigapixels) with only 5 GB of GPU VRAM at a batch size of 16.

****

Conditional Antibody Design as 3D Equivariant Graph Translation
Xiangzhe Kong,Wenbing Huang,Yang Liu

Antibody design is valuable for therapeutic usage and biological research. Existing deep-learning-based methods encounter several key issues: 1) incomplete context for Complementarity-Determining Regions (CDRs) generation; 2) incapability of capturing the entire 3D geometry of the input structure; 3) inefficient prediction of the CDR sequences in an autoregressive manner. In this paper, we propose Multi-channel Equivariant Attention Network (MEAN) to co-design 1D sequences and 3D structures of CDRs. To be specific, MEAN formulates antibody design as a conditional graph translation problem by importing extra components including the target antigen and the light chain of the antibody. Then, MEAN resorts to E(3)-equivariant message passing along with a proposed attention mechanism to better capture the geometrical correlation between different components. Finally, it outputs both the 1D sequences and 3D structure via a multi-round progressive full-shot scheme, which enjoys more efficiency and precision against previous autoregressive approaches. Our method significantly surpasses state-of-the-art models in sequence and structure modeling, antigen-binding CDR design, and binding affini ty optimization. Specifically, the relative improvement to baselines is about 23\% in antigen-binding CDR design and 34\% for affinity optimization.

****

Robust Constrained Reinforcement Learning

Yue Wang,Fei Miao,Shaofeng Zou

Constrained reinforcement learning is to maximize the reward subject to constraints on utilities/costs. However, in practice it is often the case that the training environment is not the same as the test one, due to, e.g., modeling error, adversarial attack, non-stationarity, resulting in severe performance degradation and more importantly constraint violation in the test environment. To address this challenge, we formulate the framework of robust constrained reinforcement learning under model uncertainty, where the MDP is not fixed but lies in some uncertainty set. The goal is two fold: 1) to guarantee that constraints on utilities/costs are satisfied for all MDPs in the uncertainty set, and 2) to maximize the worst-case reward performance over the uncertainty set. We design a robust primal-dual approach, and further develop theoretical guarantee on its convergence, complexity and robust feasibility. We then investigate a concrete example of δ-contamination uncertainty set, design an online and model-free algorithm and theoretically characterize its sample complexity.

****

Fuzzy Alignments in Directed Acyclic Graph for Non-Autoregressive Machine Translation

Zhengrui Ma,Chenze Shao,Shangtong Gui,Min Zhang,Yang Feng

Non-autoregressive translation (NAT) reduces the decoding latency but suffers from performance degradation due to the multi-modality problem. Recently, the structure of directed acyclic graph has achieved great success in NAT, which tackles the multi-modality problem by introducing dependency between vertices. However, training it with negative log-likelihood loss implicitly requires a strict alignment between reference tokens and vertices, weakening its ability to handle multiple translation modalities. In this paper, we hold the view that all paths in the graph are fuzzily aligned with the reference sentence. We do not require the exact alignment but train the model to maximize a fuzzy alignment score between the graph and reference, which takes captured translations in all modalities into account. Extensive experiments on major WMT benchmarks show that our method substantially improves translation performance and increases prediction confidence, setting a new state of the art for NAT on the raw training data.

****

Efficient Federated Domain Translation

Zeyu Zhou,Sheikh Shams Azam,Christopher Brinton,David I. Inouye

A central theme in federated learning (FL) is the fact that client data distributions are often not independent and identically distributed (IID), which has strong implications on the training process. While most existing FL algorithms focus on the conventional non-IID setting of class imbalance or missing classes across clients, in practice, the distribution differences could be more complex, e.g., changes in class conditional (domain) distributions. In this paper, we consider this complex case in FL wherein each client has access to only one domain distribution. For tasks such as domain generalization, most existing learning algorithms require access to data from multiple clients (i.e., from multiple domains) during training, which is prohibitive in FL. To address this challenge, we propose a federated domain translation method that generates pseudodata for each client which could be useful for multiple downstream learning tasks. We empirically demonstrate that our translation model is more resource-efficient (in terms of both communication and computation) and easier to train in an FL setting than standard domain translation methods. Furthermore, we demonstrate that the learned translation model enables use of state-of-the-art domain generalization methods in a federated setting, which enhances accuracy and robustness to increases in the synchronization period compared to existing methodology.

****

Single-Stage Open-world Instance Segmentation with Cross-task Consistency Regularization

XIZHE XUE,Dongdong Yu,Lingqiao Liu,Yu Liu,Satoshi Tsutsui,Ying Li,Zehuan Yuan,Ping Song,Mike Zheng Shou

Open-World Instance Segmentation (OWIS) is an emerging research topic that aims to segment class-agnostic object instances from images. The mainstream approaches use a two-stage segmentation framework, which first locates the candidate object bounding boxes and then performs instance segmentation. In this work, we instead promote a single-stage framework for OWIS. We argue that the end-to-end training process in the single-stage framework can be more convenient for directly regularizing the localization of class-agnostic object pixels. Based on the single-stage instance segmentation framework, we propose a regularization model to predict foreground pixels and use its relation to instance segmentation to construct a cross-task consistency loss. We show that such a consistency loss could alleviate the problem of incomplete instance annotation -- a common problem in the existing OWIS datasets. We also show that the proposed loss lends itself to an effective solution to semi-supervised OWIS that could be considered an extreme case that all object annotations are absent for some images. Our extensive experiments demonstrate that the proposed method achieves impressive results in both fully-supervised and semi-supervised settings. Compared to SOTA methods, the proposed method significantly improves the AP_{100} score by 4.75\% in UVO$\rightarrow$UVO setting and 4.05\% in COCO$\rightarrow$UVO setting. In the case of semi-supervised learning, our model learned with only 30\% labeled data, even outperforms its fully-supervised counterpart with 50\% labeled data. The code will be released soon.

****

What can be learnt with wide convolutional neural networks?

Francesco Cagnetta,Alessandro Favero,Matthieu Wyart

Understanding how convolutional neural networks (CNNs) can efficiently learn high-dimensional functions remains a fundamental challenge. A popular belief is that these models harness the local and hierarchical structure of natural data such as images. Yet, we lack a quantitative understanding of how such structure affects performance, e.g. the rate of decay of the generalisation error with the number of training samples. In this paper, we study deep CNNs in the kernel regime. First, we show that the spectrum of the corresponding kernel inherits the hierarchical structure of the network, and we characterise its asymptotics. Then, we use this result together with generalisation bounds to prove that deep CNNs adapt to the spatial scale of the target function. In particular, we find that if the target function depends on low-dimensional subsets of adjacent input variables, then the rate of decay of the error is controlled by the effective dimensionality of these subsets. Conversely, if the teacher function depends on the full set of input variables, then the error rate is inversely proportional to the input dimension. We conclude by computing the rate when a deep CNN is trained on the output of another deep CNN with randomly-initialised parameters. Interestingly, we find that despite their hierarchical structure, the functions generated by deep CNNs are too rich to be efficiently learnable in high dimension.

****

3D Segmenter: 3D Transformer based Semantic Segmentation via 2D Panoramic Distillation

ZHENNAN WU,YANG LI,Yifei Huang,Lin Gu,Tatsuya Harada,Hiroyuki Sato

Recently, 2D semantic segmentation has witnessed a significant advancement thanks to the huge amount of 2D image datasets available. Therefore, in this work, we propose the first 2D-to-3D knowledge distillation strategy to enhance 3D semantic segmentation model with knowledge embedded in the latent space of powerful 2D models. Specifically, unlike standard knowledge distillation, where teacher and student models take the same data as input, we use 2D panoramas properly aligned with corresponding 3D rooms to train the teacher network and use the learned knowledge from 2D teacher to guide 3D student. To facilitate our research, we create a large-scale, fine-annotated 3D semantic segmentation benchmark, containing voxel-wise semantic labels and aligned panoramas of 5175 scenes. Based on this benchmark, we propose a 3D volumetric semantic segmentation network, which adapts Video Swin Transformer as backbone and introduces a skip connected linear deco der. Achieving a state-of-the-art performance, our 3D Segmenter is computationally efficient and only requires 3.8% of the parameters compared to the prior art. Our code and data will be released upon acceptance.

****

Towards Skilled Population Curriculum for MARL

Rundong Wang,Longtao Zheng,Wei Qiu,Bowei He,Bo An,Zinovi Rabinovich,Yujing Hu,Yingfeng Chen,Tangjie Lv,Changjie Fan

Recent advances in multi-agent reinforcement learning (MARL) allow agents to coordinate their behaviors in complex environments. However, common MARL algorithms still suffer from scalability and sparse reward issues. One promising approach to resolve them is automated curriculum learning (ACL), where a student (curriculum learner) train on tasks of increasing difficulty controlled by a teacher (curriculum generator). Unfortunately, in spite of its success, ACL’s applicability is restricted due to: (1) lack of a general student framework to deal with the varying number of agents across tasks and the sparse reward problem, and (2) the non-stationarity in the teacher’s task due to the ever-changing student strategies. As a remedy for ACL, we introduce a novel automatic curriculum learning framework, Skilled Population Curriculum (SPC), adapting curriculum learning to multi-agent coordination. To be specific, we endow the student with population-invariant communication and a hierarchical skill set. Thus, the student can learn cooperation and behavior skills from distinct tasks with a varying number of agents. In addition, we model the teacher as a contextual bandit conditioned by student policies. As a result, a team of agents can change its size while retaining previously acquired skills. We also analyze the inherent non-stationarity of this multi-agent automatic curriculum teaching problem, and provide a corresponding regret bound. Empirical results show that our method improves scalability, sample efficiency, and generalization in multiple MARL environments. The source code and the video can be found at https://sites.google.com/view/marl-spc/.

****

Clifford Neural Layers for PDE Modeling

Johannes Brandstetter,Rianne van den Berg,Max Welling,Jayesh K Gupta

Partial differential equations (PDEs) see widespread use in sciences and engineering to describe simulation of physical processes as scalar and vector fields interacting and coevolving over time. Due to the computationally expensive nature of their standard solution methods, neural PDE surrogates have become an active research topic to accelerate these simulations. However, current methods do not explicitly take into account the relationship between different fields and their internal components, which are often correlated. Viewing the time evolution of such correlated fields through the lens of multivector fields allows us to overcome these limitations. Multivector fields consist of scalar, vector, as well as higher-order components, such as bivectors and trivectors. Their algebraic properties, such as multiplication, addition and other arithmetic operations can be described by Clifford algebras. To our knowledge, this paper presents the first usage of such multivector representations together with Clifford convolutions and Clifford Fourier transforms in the context of deep learning. The resulting Clifford neural layers are universally applicable and will find direct use in the areas of fluid dynamics, weather forecasting, and the modeling of physical systems in general. We empirically evaluate the benefit of Clifford neural layers by replacing convolution and Fourier operations in common neural PDE surrogates by their Clifford counterparts on 2D Navier-Stokes and weather modeling tasks, as well as 3D Maxwell equations. For similar parameter count, Clifford neural layers consistently improve generalization capabilities of the tested neural PDE surrogates.

****

GOOD: Exploring geometric cues for detecting objects in an open world

Haiwen Huang,Andreas Geiger,Dan Zhang

We address the task of open-world class-agnostic object detection, i.e., detecting every object in an image by learning from a limited number of base object classes. State-of-the-art RGB-based models suffer from overfitting the training classes and often fail at detecting novel-looking objects. This is because RGB-base d models primarily rely on appearance similarity to detect novel objects and are also prone to overfitting short-cut cues such as textures and discriminative parts. To address these shortcomings of RGB-based object detectors, we propose incorporating geometric cues such as depth and normals, predicted by general-purpose monocular estimators. Specifically, we use the geometric cues to train an object proposal network for pseudo-labeling unannotated novel objects in the training set. Our resulting Geometry-guided Open-world Object Detector (GOOD) significantly improves detection recall for novel object categories and already performs well with only a few training classes. Using a single ``person'' class for training on the COCO dataset, GOOD surpasses SOTA methods by 5.0% AR@100, a relative improvement of 24%. The code has been made available at https://github.com/autonomousvision/good.

****

Bringing Saccades and Fixations into Self-supervised Video Representation Learning

Qiuxia LAI,Ailing Zeng,Ye Wang,Lihong Cao,Qiang Xu

In this paper, we propose a self-supervised video representation learning (video SSL) method by taking inspiration from cognitive science and neuroscience on human visual perception. Different from previous methods that mainly start from the inherent properties of videos, we argue that humans learn to perceive the world through the self-awareness of the semantic change or consistency in the input stimuli in the absence of labels, accompanied by representation reorganization during the post-learning rest periods. To this end, we first exploit the presence of saccades as an indicator of semantic change in a contrastive learning framework to mimic the self-awareness in human representation learning, where the saccades are generated without eye-tracking data. Second, we model the semantic consistency by minimizing the prediction error between the predicted and the true state of another time point during a fixation. Third, we later incorporate prototypical contrastive learning to reorganize the learned representations such that perceptually similar representations would be associated closer. Compared to previous counterparts, our method can capture finer-grained semantics from video instances, and the associations among similar ones are further strengthened. Experiments show that the proposed bio-inspired video SSL method significantly improves the Top-1 video retrieval accuracy on UCF101 and achieves superior performance on downstream tasks such as action recognition under comparable settings.

****

Improve learning combining crowdsourced labels by weighting Areas Under the Margin

Tanguy Lefort,Benjamin Charlier,Alexis Joly,Joseph Salmon

In supervised learning -- for instance in image classification -- modern massive datasets are commonly labelled by a crowd of workers. The obtained labels in this crowdsourcing setting are then aggregated for training. The aggregation step generally leverages a per worker trust score. Yet, such worker-centric approaches discard each task ambiguity. Some intrinsically ambiguous tasks might even fool expert workers, which could eventually be harmful for the learning step. In a standard supervised learning setting -- with one label per task and balanced classes -- the Area Under the Margin (AUM) statistic is tailored to identify mislabeled data. We adapt the AUM to identify ambiguous tasks in crowdsourced learning scenarios, introducing the Weighted AUM (WAUM). The WAUM is an average of AUMs weighted by worker and task dependent scores. We show that the WAUM can help discarding ambiguous tasks from the training set, leading to better generalization or calibration performance. We report improvements with respect to feature-blind aggregation strategies both for simulated settings and for the CIFAR-10H crowdsourced dataset.

****

Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task

Kenneth Li,Aspen K Hopkins,David Bau,Fernanda Viégas,Hanspeter Pfister,Martin Wattenberg

Language models show a surprising range of capabilities, but the source of their apparent competence is unclear. Do these networks just memorize a collection of surface statistics, or do they rely on internal representations of the process that generates the sequences they see? We investigate this question by applying a variant of the GPT model to the task of predicting legal moves in a simple board game, Othello. Although the network has no a priori knowledge of the game or its rules, we uncover evidence of an emergent nonlinear internal representation of the board state. Interventional experiments indicate this representation can be used to control the output of the network and create "latent saliency maps" that can help explain predictions in human terms.

****

Programmatically Grounded, Compositionally Generalizable Robotic Manipulation

Renhao Wang,Jiayuan Mao,Joy Hsu,Hang Zhao,Jiajun Wu,Yang Gao

Robots operating in the real world require both rich manipulation skills as well as the ability to semantically reason about when to apply those skills. Towards this goal, recent works have integrated semantic representations from large-scale pretrained vision-language (VL) models into manipulation models, imparting them with more general reasoning capabilities. However, we show that the conventional {\it pretraining-finetuning} pipeline for integrating such representations entangles the learning of domain-specific action information and domain-general visual information, leading to less data-efficient training and poor generalization to unseen objects and tasks. To this end, we propose \ours, a {\it modular} approach to better leverage pretrained VL models by exploiting the syntactic and semantic structures of language instructions. Our framework uses a semantic parser to recover an executable program, composed of functional modules grounded on vision and action across different modalities. Each functional module is realized as a combination of deterministic computation and learnable neural networks. Program execution produces parameters to general manipulation primitives for a robotic end-effector. The entire modular network can be trained with end-to-end imitation learning objectives. Experiments show that our model successfully disentangles action and perception, translating to improved zero-shot and compositional generalization in a variety of manipulation behaviors. Project webpage at: \url{https://progport.github.io}.

****

ObPose: Leveraging Pose for Object-Centric Scene Inference and Generation in 3D

Yizhe Wu,Oiwi Parker Jones,Ingmar Posner

We present ObPose, an unsupervised object-centric inference and generation model which learns 3D-structured latent representations from RGB-D scenes. Inspired by prior art in 2D representation learning, ObPose considers a factorised latent space, separately encoding object location (where) and appearance (what). ObPose further leverages an object's pose (i.e. location and orientation), defined via a minimum volume principle, as a novel inductive bias for learning the where component. To achieve this, we propose an efficient, voxelised approximation approach to recover the object shape directly from a neural radiance field (NeRF). As a consequence, ObPose models each scene as a composition of NeRFs, richly representing individual objects. To evaluate the quality of the learned representations, ObPose is evaluated quantitatively on the YCB and CLEVR datatasets for unsupervised scene segmentation, outperforming the current state-of-the-art in 3D scene inference (ObSuRF) by a significant margin. Generative results provide qualitative demonstration that the same ObPose model can both generate novel scenes and flexibly edit the objects in them. These capacities again reflect the quality of the learned latents and the benefits of disentangling the where and what components of a scene. Key design choices made in the ObPose encoder are validated with ablations.

****

FedCL: Critical Learning Periods-aware Adaptive Client Selection in Federated Learning

Gang Yan,Hao Wang,Xu Yuan,Jian Li

Federated learning (FL) is a distributed optimization paradigm that learns from data samples distributed across a number of clients. Adaptive client selection that is cognizant of the training progress of clients has become a major trend t o improve FL efficiency but not yet well-understood. Most existing FL methods such as FedAvg and its state-of-the-art variants implicitly assume that all learning phases during the FL training process are equally important. Unfortunately, this assumption has been revealed to be invalid due to recent findings on critical learning (CL) periods, in which small gradient errors may lead to an irrecoverable deficiency on final test accuracy. In this paper, we develop FedCL, a CL periods-aware FL framework to reveal that adaptively augmenting exiting FL methods with CL periods, the resultant performance is significantly improved when the client selection is guided by the discovered CL periods. Experiments based on various machine learning models and datasets validate that the proposed FedCL framework consistently achieves an improved model accuracy while maintains comparable or even better communication efficiency as compared to state-of-the-art methods, demonstrating a promising and easily adopted method for tackling the heterogeneity of FL training.

****

TabCaps: A Capsule Neural Network for Tabular Data Classification with BoW Routing

Jintai Chen,KuanLun Liao,Yanwen Fang,Danny Chen,Jian Wu

Records in a table are represented by a collection of heterogeneous scalar features. Previous work often made predictions for records in a paradigm that processed each feature as an operating unit, which requires to well cope with the heterogeneity. In this paper, we propose to encapsulate all feature values of a record into vectorial features and process them collectively rather than have to deal with individual ones, which directly captures the representations at the data level and benefits robust performances. Specifically, we adopt the concept of "capsules" to organize features into vectorial features, and devise a novel capsule neural network called "TabCaps" to process the vectorial features for classification. In TabCaps, a record is encoded into several vectorial features by some optimizable multivariate Gaussian kernels in the primary capsule layer, where each vectorial feature represents a specific "profile" of the input record and is transformed into senior capsule layer under the guidance of a new straightforward routing algorithm. The design of routing algorithm is motivated by the Bag-of-Words (BoW) model, which performs capsule feature grouping straightforwardly and efficiently, in lieu of the computationally complex clustering of previous routing algorithms. Comprehensive experiments show that TabCaps achieves competitive and robust performances in tabular data classification tasks.

****

Learning Instance-Solution Operator For Optimal Control

Mingquan Feng,Zhijie Chen,Junchi Yan

Optimal control problems (OCPs) aim at finding a control function for a dynamical system such that a cost functional is optimized. These problems are central to physical system research in both academia and industry. In this paper, we propose a novel instance-solution operator learning perspective, which solves OCPs in a one-shot manner with no dependence on the explicit expression of dynamics or iterative optimization processes. The design is in principle endowed with substantial speedup in running time, and the model reusability is guaranteed by high-quality in- and out-of-distribution generalization. We theoretically validate the perspective by presenting the approximation bounds for the instance-solution operator learning. Extensive experiments on 6 physical systems verify the effectiveness and efficiency of our approach. The source code will be made publicly available.

****

Decentralized Online Bandit Optimization on Directed Graphs with Regret Bounds

Johan Östman,Ather Gattami,Daniel Gillblad

We consider a decentralized multiplayer game, played over T rounds, with a leader-follower hierarchy described by a directed acyclic graph. For each round, the graph structure dictates the order of the players and how players observe the actions of one another. By the end of each round, all players receive a joint bandit-reward based on their joint action that is used to update the player strate gies towards the goal of minimizing the joint pseudo-regret. We present a learning algorithm inspired by the single-player multi-armed bandit problem and show that it achieves sub-linear joint pseudo-regret in the number of rounds for both adversarial and stochastic bandit rewards. Furthermore, we quantify the cost incurred due to the decentralized nature of our problem compared to the centralized setting.

****

BAMBI: Vertical Federated Bilevel Optimization with Privacy-Preserving and Computation Efficiency

Qingsong Zhang,Fengxiang He,Jindong Gu,Bin Gu,Cheng Deng,Heng Huang,Dacheng Tao

Vertical federated learning (VFL) has shown promising in meeting the vast demands of multi-party privacy-preserving learning. However, existing VFL methods are not applicable to popular machine learning tasks falling under bilevel programming, such as hyper-representation learning and hyperparameter tuning. A desirable solution is adopting bilevel optimization (BO) into VFL, but on-shelf BO methods are shackled by the difficulty in computing the hypergradients with privacy-preserving and computation-efficient under the setting of VFL. To address this challenge, this paper proposes a stochastic Bilevel optimizAtion Method with a desirable JacoBian estImator (BAMBI), which constructs a novel zeroth-order (ZO) estimator to locally approximate the Jacobian matrix. This approximation enables BAMBI to compute the hypergradients in a privacy-preserving and computation-efficient manner. We prove that BAMBI convergences in the rate of $\mathcal{O}(1/\sqrt{K})$ (K is the total number of the upper-level iterations) under the nonconvex-strongly-convex setting which covers most practical scenarios. This convergence rate is comparable with the algorithms without ZO estimator, which justifies our advantage in privacy preservation without sacrifice in convergence rate. Moreover, we design a BAMBI-DP method for further mitigating the concerns on label privacy by leveraging the differential privacy (DP) technique. Extensive experiments fully support our algorithms. The code will be released publicly. To our best knowledge, this is the first work on the bilevel optimization under the setting of VFL.

****

Revitalize Region Feature for Democratizing Video-language Pre-training of Retrieval

Guanyu Cai,Yixiao Ge,Binjie Zhang,Jinpeng Wang,Rui Yan,Xudong Lin,Ying Shan,Lianghua He,Xiaohu Qie,Jianping Wu,Mike Zheng Shou

Recent dominant methods for video-language pre-training (VLP) learn transferable representations from the raw pixels in an end-to-end manner to achieve advanced performance on downstream video-language retrieval. Despite the impressive results, VLP research becomes extremely expensive with the need for massive data and a long training time, preventing further explorations. In this work, we revitalize region features of sparsely sampled video clips to significantly reduce both spatial and temporal visual redundancy towards democratizing VLP research at the same time achieving state-of-the-art results. Specifically, to fully explore the potential of region features, we introduce a novel bidirectional region-word alignment regularization that properly optimizes the fine-grained relations between regions and certain words in sentences, eliminating the domain/modality disconnections between pre-extracted region features and text. Extensive results of downstream video-language retrieval tasks on four datasets demonstrate the superiority of our method on both effectiveness and efficiency, e.g., our method achieves competing results with 80% fewer data and 85% less pre-training time compared to the most efficient VLP method so far.

****

MESSAGENET: MESSAGE CLASSIFICATION USING NATURAL LANGUAGE PROCESSING AND META-DATA

Adar Kahana,Oren Elisha

In this paper we propose a new Deep Learning (DL) approach for message classification. Our method
is based on the state-of-the-art Natural Language Processing (NLP) building blocks, combined

with a novel technique for infusing the meta-data input that is typically available in messages such as the sender information, timestamps, attached image, audio, affiliations, and more. As we demonstrate throughout the paper, going beyond the mere text by leveraging all available channels in the message, could yield an improved representation and higher classification accuracy. To achieve message representation, each type of input is processed in a dedicated block in the neural network architecture that is suitable for the data type. Such an implementation enables training all blocks together simultaneously, and forming cross channels features in the network. We show in the Experiments Section that in some cases, message’s meta-data holds an additional information that cannot be extracted just from the text, and when using this information we achieve better performance.
Furthermore, we demonstrate that our multi-modality block approach outperforms other approaches for injecting the meta data to the the text classifier.
****

Universal approximation and model compression for radial neural networks

Iordan Ganev,Twan van Laarhoven,Robin Walters

We introduce a class of fully-connected neural networks whose activation functions, rather than being pointwise, rescale feature vectors by a function depending only on their norm. We call such networks radial neural networks, extending previous work on rotation equivariant networks that considers rescaling activations in less generality. We prove universal approximation theorems for radial neural networks, including in the more difficult cases of bounded widths and unbounded domains. Our proof techniques are novel, distinct from those in the pointwise case. Additionally, radial neural networks exhibit a rich group of orthogonal change-of-basis symmetries on the vector space of trainable parameters. Factoring out these symmetries leads to a practical lossless model compression algorithm. Optimization of the compressed model by gradient descent is equivalent to projected gradient descent for the full model.
****

Momentum Diminishes the Effect of Spectral Bias in Physics-Informed Neural Networks

Ghazal Farhani,Alexander Kazachek,Boyu Wang

Physics-informed neural network (PINN) algorithms have shown promising results in solving a wide range of problems involving partial differential equations (PDEs). However, even the simplest PDEs, often fail to converge to desirable solutions when the target function contains high-frequency modes, due to a phenomenon known as spectral bias. In the present work, we exploit neural tangent kernels (NTKs) to investigate the training dynamics of PINNs evolving under stochastic gradient descent with momentum (SGDM). This demonstrates SGDM significantly reduces the effect of spectral bias. We have also examined why training a model via the Adam optimizer can accelerate the convergence while reducing the spectral bias. Moreover, our numerical experiments have confirmed that wide-enough networks using SGDM or Adam still converge to desirable solutions, even in the presence of high-frequency features.
****

MULTILEVEL XAI: VISUAL AND LINGUISTIC BONDED EXPLANATIONS

Halil Ibrahim Aysel,Xiaohao Cai,Adam Prugel-Bennett

Applications of deep neural networks are booming in more and more fields but lack transparency due to their black-box nature. Explainable Artificial Intelligence (XAI) is therefore of paramount importance, where strategies are proposed to understand how these black-box models function. The research so far mainly focuses on producing, for example, class-wise saliency maps, highlighting parts of a g

iven image that affect the prediction the most. However, this way does not fully represent the way humans explain their reasoning and, awkwardly, validating these maps is quite complex and generally requires subjective interpretation. In this article, we conduct XAI differently by proposing a new XAI methodology in a multilevel (i.e., visual and linguistic) manner. By leveraging the interplay between the learned representations, i.e., image features and linguistic attributes, the proposed approach can provide salient attributes and attribute-wise saliency maps, which are far more intuitive than the class-wise maps, without requiring per-image ground-truth human explanations. It introduces self-interpretable attributes to overcome the current limitations in XAI and bring the XAI towards human-like level. The proposed architecture is simple in use and can reach surprisingly good performance in both prediction and explainability for deep neural networks thanks to the low-cost per-class attributes.
****

An Exact Poly-Time Membership-Queries Algorithm for Extracting a Three-Layer ReLU Network

Amit Daniely,Elad Granot

We consider the natural problem of learning a ReLU network from queries, which was recently remotivated by model extraction attacks. In this work, we present a polynomial-time algorithm that can learn a depth-two ReLU network from queries under mild general position assumptions. We also present a polynomial-time algorithm that, under mild general position assumptions, can learn a rich class of depth-three ReLU networks from queries. For instance, it can learn most networks where the number of first layer neurons is smaller than the dimension and the number of second layer neurons.

These two results substantially improve state-of-the-art: Until our work, polynomial-time algorithms were only shown to learn from queries depth-two networks under the assumption that either the underlying distribution is Gaussian (Chen et al. (2021)) or that the weights matrix rows are linearly independent (Milli et al. (2019)). For depth three or more, there were no known poly-time results.
****

Neural Discrete Reinforcement Learning

Yazhe Niu,Yuan Pu,Chuming Li,Zhenjie Yang,Hongsheng Li,Yu Liu

Designing effective action spaces for complex environments is a fundamental and challenging problem in reinforcement learning (RL).

Some recent works have revealed that naive RL algorithms utilizing well-designed handcrafted discrete action spaces can achieve promising results even when dealing with high-dimensional continuous or hybrid decision-making problems. However, elaborately designing such action spaces requires comprehensive domain knowledge.

In this paper, we systemically analyze the advantages of discretization for different action spaces and then propose a unified framework, Neural Discrete Reinforcement Learning (NDRL), to automatically learn how to effectively discretize almost arbitrary action spaces.

Specifically, we propose the Action Discretization Variational AutoEncoder (AD-VAE), an action representation learning method that can learn compact latent action spaces while maintain the essential properties of original environments, such as boundary actions and the relationship between different action dimensions. Moreover, we uncover a key issue that parallel optimization of the AD-VAE and online RL agents is often unstable. To address it, we further design several techniques to adapt RL agents to learned action representations, including latent action remapping and ensemble Q-learning. Quantitative experiments and visualization results demonstrate the efficiency and stability of our proposed framework for complex action spaces in various environments.
****

CAB: Comprehensive Attention Benchmarking on Long Sequence Modeling

Jun Zhang,Shuyang Jiang,Jiangtao Feng,Lin Zheng,Lingpeng Kong

Transformer has achieved remarkable success in language, image, and speech processing. Recently, various efficient attention architectures have been proposed to improve transformer's efficiency while largely preserving its efficacy, especially in modeling long sequences. A widely-used benchmark to test these efficient methods' capability on long-range modeling is Long Range Arena (LRA). However, LRA only focuses on the standard bidirectional (or noncausal) self attention, and completely ignores cross attentions and unidirectional (or causal) attentions, which are equally important to downstream applications. Although designing cross and causal variants of an attention method is straightforward for vanilla attention, it is often challenging for efficient attentions with subquadratic time and memory complexity. In this paper, we propose Comprehensive Attention Benchmark (CAB) under a fine-grained attention taxonomy with four distinguishable attention patterns, namely, noncausal self, causal self, noncausal cross, and causal cross attentions. CAB collects seven real-world tasks from different research areas to evaluate efficient attentions under the four attention patterns. Among these tasks, CAB validates efficient attentions in eight backbone networks to show their generalization across neural architectures. We conduct exhaustive experiments to benchmark the performances of nine widely-used efficient attention architectures designed with different philosophies on CAB. Extensive experimental results also shed light on the fundamental problems of efficient attentions, such as efficiency length against vanilla attention, performance consistency across attention patterns, the benefit of attention mechanisms, and interpolation/extrapolation on long-context language modeling.

****

Formal Conceptual Views in Neural Networks

Johannes Hirth,Tom Hanika

Explaining neural network models is a challenging task that remains unsolved in its entirety to this day. This is especially true for high dimensional and complex data. With the present work, we introduce two notions for conceptual views of a neural network, specifically a many-valued and a symbolic view. Both provide novel analysis methods to enable a human AI analyst to grasp deeper insights into the knowledge that is captured by the neurons of a network. We test the conceptual expressivity of our novel views through different experiments on the ImageNet and Fruit-360 data sets. Furthermore, we show to which extent the views allow to quantify the conceptual similarity of different learning architectures. Finally, we demonstrate how conceptual views can be applied for abductive learning of human comprehensible rules from neurons. In summary, with our work, we contribute to the most relevant task of globally explaining neural networks models.

****

Towards Understanding and Mitigating Dimensional Collapse in Heterogeneous Federated Learning

Yujun Shi,Jian Liang,Wenqing Zhang,Vincent Tan,Song Bai

Federated learning aims to train models collaboratively across different clients without sharing data for privacy considerations. However, one major challenge for this learning paradigm is the data heterogeneity problem, which refers to the discrepancies between the local data distributions among various clients. To tackle this problem, we first study how data heterogeneity affects the representations of the globally aggregated models. Interestingly, we find that heterogeneous data results in the global model suffering from severe dimensional collapse, in which representations tend to reside in a lower-dimensional space instead of the ambient space. Moreover, we observe a similar phenomenon on models locally trained on each client and deduce that the dimensional collapse on the global model is inherited from local models. In addition, we theoretically analyze the gradient flow dynamics to shed light on how data heterogeneity result in dimensional collapse for local models. To remedy this problem caused by the data heterogeneity, we propose FedDecorr, a novel method that can effectively mitigate dimensional collapse in federated learning. Specifically, FedDecorr applies a regularization term during local training that encourages different dimensions of representations to be uncorrelated. FedDecorr, which is implementation-friendly and computationally-efficient, yields consistent improvements over baselines on standard benchmark datasets. Code: https://github.com/bytedance/FedDecorr.

****

A New Paradigm for Federated Structure Non-IID Subgraph Learning

Xunkai Li,Wentao Zhang,Rong-Hua Li,Yulin Zhao,Yinlin Zhu,Guoren Wang

Federated graph learning (FGL), a distributed training framework for graph neural networks (GNNs) has attracted much attention for breaking the centralized machine learning assumptions. Despite its effectiveness, the differences in data collection perspectives and quality lead to the challenges of heterogeneity, especially the domain-specific graph is partitioned into subgraphs in different institutions. However, existing FGL methods implement graph data augmentation or personalization with community split which follows the cluster homogeneity assumptions. Hence we investigate the above issues and suggest that subgraph heterogeneity is essentially the structure variations. From the observations on FGL, we first define the structure non-independent identical distribution (Non-IID) problem, which presents covariant shift challenges among client-wise subgraphs. Meanwhile, we propose a new paradigm for general federated data settings called Adaptive Federated Graph Learning (AdaFGL). The motivation behind it is to implement adaptive propagation mechanisms based on federated global knowledge and non-params label propagation. We conduct extensive experiments with community split and structure Non-IID settings, our approach achieves state-of-the-art performance on five benchmark datasets.

****

SketchKnitter: Vectorized Sketch Generation with Diffusion Models

Qiang Wang,Haoge Deng,Yonggang Qi,Da Li,Yi-Zhe Song

We show vectorized sketch generation can be identified as a reversal of the stroke deformation process. This relationship was established by means of a diffusion model that learns data distributions over the stroke-point locations and pen states of real human sketches. Given randomly scattered stroke-points, sketch generation becomes a process of deformation-based denoising, where the generator rectifies positions of stroke points at each timestep to converge at a recognizable sketch. A key innovation was to embed recognizability into the reverse time diffusion process. It was observed that the estimated noise during the reversal process is strongly correlated with sketch classification accuracy. An auxiliary recurrent neural network (RNN) was consequently used to quantify recognizability during data sampling. It follows that, based on the recognizability scores, a sampling shortcut function can also be devised that renders better quality sketches with fewer sampling steps. Finally it is shown that the model can be easily extended to a conditional generation framework, where given incomplete and unfaithful sketches, it yields one that is more visually appealing and with higher recognizability.

****

Evidential Uncertainty and Diversity Guided Active Learning for Scene Graph Generation

Shuzhou Sun,Shuaifeng Zhi,Janne Heikkilä,Li Liu

Scene Graph Generation (SGG) has already shown its great potential in various downstream tasks, but it comes at the price of a prohibitively expensive annotation process. To reduce the annotation cost, we propose using Active Learning (AL) for sampling the most informative data. However, directly porting current AL methods to the SGG task poses the following challenges: 1) unreliable uncertainty estimates, and 2) data bias problems. To deal with these challenges, we propose EDAL (\textbf{E}vidential Uncertainty and \textbf{D}iversity Guided Deep \textbf{A}ctive \textbf{L}earning), a novel AL framework tailored for the SGG task. For challenge 1), we start with Evidential Deep Learning (EDL) coupled with a global relationship mining approach to estimate uncertainty, which can effectively overcome the perturbations of open-set relationships and background-relationships to obtain reliable uncertainty estimates. To address challenge 2), we seek the diversity-based method and design the Context Blocking Module (CBM) and Image Blocking Module (IBM) to alleviate context-level bias and image-level bias, respectively. Experiments show that our AL framework can approach the performance of a fully supervised SGG model with only about 10% annotation cost. Furthermore, our ablation studies indicate that introducing AL into the SGG will face many challenges not observed in other vision tasks that are successfully overcome by our

new modules.
****

Test-time recalibration of conformal predictors under distribution shift based on unlabeled examples

Fatih Furkan Yilmaz,Reinhard Heckel

Modern image classifiers achieve high predictive accuracy, but the predictions typically come without reliable uncertainty estimates. Conformal prediction algorithms provide uncertainty estimates by predicting a set of classes based on the probability estimates of the classifier (for example, the softmax scores). To provide such sets, conformal prediction algorithms often rely on estimating a cutoff threshold for the probability estimates, and this threshold is chosen based on a calibration set. Conformal prediction methods guarantee reliability only when the calibration set is from the same distribution as the test set. Therefore, the methods need to be recalibrated for new distributions. However, in practice, labeled data from new distributions is rarely available, making calibration infeasible. In this work, we consider the problem of predicting the cutoff threshold for a new distribution only based on unlabeled examples. While it is impossible in general to guarantee reliability when calibrating based on unlabeled examples, we show that our method provides excellent uncertainty estimates under natural distribution shifts.

****

TabDDPM: Modelling Tabular Data with Diffusion Models

Akim Kotelnikov,Dmitry Baranchuk,Ivan Rubachev,Artem Babenko

Denoising diffusion probabilistic models are currently becoming the leading paradigm of generative modeling for many important data modalities. Being the most prevalent in the computer vision community, diffusion models have also recently gained some attention for other domains, including speech, NLP, and graph-like data. In this work, we investigate if the framework of diffusion models can be advantageous for general tabular problems, where datapoints are typically represented by vectors of heterogeneous features. The inherent heterogeneity of tabular data makes it quite challenging for accurate modeling, since the individual features can be of completely different nature, i.e., some of them can be continuous and some of them can be discrete. To address such data types, we introduce TabDDPM --- a diffusion model that can be universally applied to any tabular dataset and handles any types of features. We extensively evaluate TabDDPM on a wide set of benchmarks and demonstrate its superiority over existing GAN/VAE alternatives, which is consistent with the advantage of diffusion models in other fields. Additionally, we show that TabDDPM can be successfully used in privacy-oriented setups, where the original datapoints cannot be shared.

****

BED: Boundary-Enhanced Decoder for Chinese Word Segmentation

Shiting Xu,Dongge Tang,Weiwei Jiang,Qing Yang

Chinese Word Segmentation (CWS) is an essential fundamental step in the Chinese NLP processing pipeline. In recent years, with the development of deep learning and pre-training language models, many CWS models based on pre-training models, e.g., BERT and Roberta, have been proposed, and the performance of CWS models has been dramatically improved. However, CWS remains an open problem that deserves further study, such as the poor effect on OOV words. To our knowledge, the current proposed CWS approaches mainly focus on optimizing the encoder part of the model, such as incorporating more word information into the encoder or doing pre-training related to the CWS task, etc. And there is no attempt to improve the decoder's performance in the CWS model. This paper proposes an optimized decoder for the CWS model called Boundary-Enhanced Decoder (BED). It could bring 0.05% and 0.69% improvement on Average-F1 and OOV Average-F1 on four benchmark datasets when using a model with a BERT encoder and softmax standard decoder. We also publish our implementation of BED.

****

Tailoring Language Generation Models under Total Variation Distance

Haozhe Ji,Pei Ke,Zhipeng Hu,Rongsheng Zhang,Minlie Huang

The standard paradigm of neural language generation adopts maximum likelihood es

timation (MLE) as the optimizing method. From a distributional view, MLE in fact minimizes the Kullback-Leibler divergence (KLD) between the distribution of the real data and that of the model. However, this approach forces the model to distribute non-zero (sometimes large) probability mass to all training samples regardless of their quality. Moreover, in the attempt to cover the low-probability regions in the data distribution, the model systematically overestimates the probability of corrupted text sequences, which we conjecture is one of the main reasons for text degeneration during autoregressive decoding. To remedy this problem, we leverage the total variation distance (TVD) with its robustness to outliers, and develop practical bounds to apply it to language generation. Then, we introduce the TaiLr objective that balances the tradeoff of estimating TVD. Intuitively, TaiLr downweights real data samples that have low model probabilities with tunable penalization intensity. Experimental results show that our method alleviates the overestimation of degenerated sequences without sacrificing diversity and improves generation quality on a wide range of text generation tasks.
****
SeqSHAP: Subsequence Level Shapley Value Explanations for Sequential Predictions
Guanyu Jiang,Fuzhen Zhuang,Bowen Song,Jiani Li,Ying Sun,Yongchun Zhu,Tianyi Zhang,Weiqiang Wang,deqing wang
With the increasing demands of interpretability in real-world applications, various methods for explainable artificial intelligence (XAI) have been proposed. However, most of them overlook the interpretability in sequential scenarios, which have a wide range of applications, e.g., online transactions and sequential recommendations. In this paper, we propose a Shapley value based explainer named SeqSHAP to explain the model predictions in sequential scenarios. Compared to existing methods, SeqSHAP provides more intuitive explanations at a subsequence level, which explicitly models the effect of contextual information among the related elements in a sequence. We propose to calculate subsequence-level feature attributions instead of element-wise attributions to utilize the information embedded in sequence structure, and provide a distribution-based segmentation method to obtain reasonable subsequences. Extensive experiments on two online transaction datasets from a real-world e-commerce platform show that the proposed method could provide valid and reliable explanations for sequential predictions.
****
Newton Losses: Efficiently Including Second-Order Information into Gradient Descent
Felix Petersen,Christian Borgelt,Tobias Sutter,Hilde Kuehne,Oliver Deussen
We present Newton losses, a method for incorporating second-order information of losses by approximating them with quadratic functions. The presented method is applied only to the loss function and allows training the neural network with gradient descent. As loss functions are usually substantially cheaper to compute than the neural network, Newton losses can be used at a relatively small additional cost. We find that they yield superior performance, especially when applied to non-convex and hard-to-optimize loss functions such as algorithmic losses, which have been popularized in recent research.
****
Anisotropic Message Passing: Graph Neural Networks with Directional and Long-Range Interactions
Moritz Thürlemann,Sereina Riniker
Graph neural networks have shown great potential for the description of a variety of chemical systems.
However, standard message passing does not explicitly account for long-range and directional interactions, for instance due to electrostatics.
In this work, an anisotropic state based on Cartesian multipoles is proposed as an addition to the existing hidden features.
With the anisotropic state, message passing can be modified to explicitly account for directional interactions.
Compared to existing models, this modification results in relatively little additional computational cost.
Most importantly, the proposed formalism offers as a distinct advantage the seam

less integration of (1) anisotropic long-range interactions, (2) interactions with surrounding fields and particles that are not part of the graph, and (3) the fast multipole method.
As an exemplary use case, the application to quantum mechanics/molecular mechanics (QM/MM) systems is demonstrated.

****

Learn Low-dimensional Shortest-path Representation of Large-scale and Complex Graphs
Haoyu Wang,Chun Yuan,Lei Li,Jiahui Jin
Estimation of shortest-path (SP) distance lies at the heart of network analysis tasks. Along with the rapid emergence of large-scale and complex graphs, approximate SP-representing algorithms that transform a graph into compact and low-dimensional representations are critical for fast and scalable online analysis. Among different approaches, learning-based representation methods have made a breakthrough both in response time and accuracy. Several competitive works in learning-based methods heuristically leverage truncated random walk and optimization on the arbitrary linkage for SP representation learning. However, they have limitations on both exploration range and distance preservation. We propose in this paper an efficient and interpretable SP representation method called Betweenness Centrality-based Distance Resampling (BCDR). First, we prove that betweenness centrality-based random walk can occupy a wider exploration range of distance due to its awareness of high-order path structures. Second, we leverage distance resampling to simulate random shortest paths from original paths and prove that the optimization on such shortest paths preserves distance relations via implicitly decomposing SP distance-based similarity matrix. BCDR yields an average improvement of 25% accuracy and 25-30% query speed, compared to all existing approximate methods when evaluated on a broad class of real-world and synthetic graphs with diverse sizes and structures.

****

SYNC: SAFETY-AWARE NEURAL CONTROL FOR STABILIZING STOCHASTIC DELAY-DIFFERENTIAL EQUATIONS
Jingdong Zhang,Qunxi Zhu,Wei Yang,Wei Lin
Stabilization of the systems described by \textit{stochastic delay}-differential equations (SDDEs) under preset conditions is a challenging task in the control community. Here, to achieve this task, we leverage neural networks to learn control policies using the information of the controlled systems in some prescribed regions. Specifically, two learned control policies, i.e., the neural deterministic controller (NDC) and the neural stochastic controller (NSC), work effectively in the learning procedures that rely on, respectively, the well-known LaSalle-type theorem and the newly-established theorem for guaranteeing the stochastic stability in SDDEs. We theoretically investigate the performance of the proposed controllers in terms of convergence time and energy cost. More practically and significantly, we improve our learned control policies through considering the situation where the controlled trajectories only evolve in some specific safety set. {\color{black} The practical validity of such control policies restricted in safety set is attributed to the theory that we further develop for safety and stability guarantees in SDDEs using the stochastic control barrier function and the spatial discretization}. We call this control as SYNC (\textbf{S}afet\textbf{Y}-aware \textbf{N}eural \textbf{C}ontrol). The efficacy of all the articulated control policies, including the SYNC, is demonstrated systematically by using representative control problems.

****

Byzantine-robust Decentralized Learning via ClippedGossip
Lie He,Sai Praneeth Karimireddy,Martin Jaggi
In this paper, we study the challenging task of Byzantine-robust decentralized training on arbitrary communication graphs. Unlike federated learning where workers communicate through a server, workers in the decentralized environment can only talk to their neighbors, making it harder to reach consensus and benefit from collaborative training. To address these issues, we propose an algorithm, termed ClippedGossip, for Byzantine-robust consensus and optimization, which is the f

irst to provably converge to a $O(\delta_{\max}\zeta^2/\gamma^2)$ neighborhood of the stationary point for non-convex objectives under standard assumptions. Finally, we demonstrate the encouraging empirical performance of ClippedGossip under a large number of attacks.

****

A Model or 603 Exemplars: Towards Memory-Efficient Class-Incremental Learning

Da-Wei Zhou,Qi-Wei Wang,Han-Jia Ye,De-Chuan Zhan

Real-world applications require the classification model to adapt to new classes without forgetting old ones. Correspondingly, Class-Incremental Learning (CIL) aims to train a model with limited memory size to meet this requirement. Typical CIL methods tend to save representative exemplars from former classes to resist forgetting, while recent works find that storing models from history can substantially boost the performance. However, the stored models are not counted into the memory budget, which implicitly results in unfair comparisons. We find that when counting the model size into the total budget and comparing methods with aligned memory size, saving models do not consistently work, especially for the case with limited memory budgets. As a result, we need to holistically evaluate different CIL methods at different memory scales and simultaneously consider accuracy and memory size for measurement. On the other hand, we dive deeply into the construction of the memory buffer for memory efficiency. By analyzing the effect of different layers in the network, we find that shallow and deep layers have different characteristics in CIL. Motivated by this, we propose a simple yet effective baseline, denoted as MEMO for Memory-efficient Expandable MOdel. MEMO extends specialized layers based on the shared generalized representations, efficiently extracting diverse representations with modest cost and maintaining representative exemplars. Extensive experiments on benchmark datasets validate MEMO's competitive performance. Code is available at: https://github.com/wangkiw/ICLR23-MEMO

****

Reinforcement learning for instance segmentation with high-level priors

Paul Hilt,Edgar Kaziakhmedov,Maedeh Zarvandi,Sourabh Bhide,Maria Leptin,Constantin Pape,Anna Kreshuk

Instance segmentation is a fundamental computer vision problem which remains challenging despite impressive recent advances due to deep learning-based methods. Given sufficient training data, fully supervised methods can yield excellent performance, but annotation of groundtruth data remains a major bottleneck, especially for biomedical applications where it has to be performed by domain experts. The amount of labels required can be drastically reduced by using rules derived from prior knowledge to guide the segmentation. However, these rules are in general not differentiable and thus cannot be used with existing methods. Here, we revoke this requirement by using stateless actor critic reinforcement learning, which enables non-differentiable rewards. We formulate the instance segmentation problem as graph partitioning and the actor critic predicts the edge weights driven by the rewards, which are based on the conformity of segmented instances to high-level priors on object shape, position or size. The experiments on toy and real data demonstrate that a good set of priors is sufficient to reach excellent performance without any direct object-level supervision.

****

Differentiable Mathematical Programming for Object-Centric Representation Learning

Adeel Pervez,Phillip Lippe,Efstratios Gavves

We propose topology-aware feature partitioning into k disjoint partitions for given scene features as a method for object-centric representation learning. To this end, we propose to use minimum s-t graph cuts as a partitioning method which is represented as a linear program. The method is topologically aware since it explicitly encodes neighborhood relationships in the image graph. To solve the graph cuts our solution relies on an efficient, scalable, and differentiable quadratic programming approximation. Optimizations specific to cut problems allow us to solve the quadratic programs and compute their gradients significantly more efficiently compared with the general quadratic programming approach. Our r

esults show that our approach is scalable and outperforms existing methods on object discovery tasks with textured scenes and objects.
****
Transformers are Sample-Efficient World Models
Vincent Micheli,Eloi Alonso,François Fleuret
Deep reinforcement learning agents are notoriously sample inefficient, which considerably limits their application to real-world problems. Recently, many model-based methods have been designed to address this issue, with learning in the imagination of a world model being one of the most prominent approaches. However, while virtually unlimited interaction with a simulated environment sounds appealing, the world model has to be accurate over extended periods of time. Motivated by the success of Transformers in sequence modeling tasks, we introduce IRIS, a data-efficient agent that learns in a world model composed of a discrete autoencoder and an autoregressive Transformer. With the equivalent of only two hours of gameplay in the Atari 100k benchmark, IRIS achieves a mean human normalized score of 1.046, and outperforms humans on 10 out of 26 games, setting a new state of the art for methods without lookahead search. To foster future research on Transformers and world models for sample-efficient reinforcement learning, we release our code and models at https://github.com/eloialonso/iris.
****
Considering Layerwise Importance in the Lottery Ticket Hypothesis
Benjamin Vandersmissen,Jose Oramas
The recently-introduced Lottery Ticket Hypothesis (LTH) posits that it is possible to extract a sparse trainable subnetwork from a dense network using iterative magnitude pruning.
By iteratively training the model, removing the connections with the lowest global weight magnitude and rewinding the remaining connections, sparse networks can be extracted that, when fully trained, reach a similar or better performance than their dense counterpart.

Intuitively, this approach of comparing connection weights globally removes a lot of context about the connection weights and their relations to other connections in their layer as the weight distributions in layers throughout the network often differ significantly.

In this paper we study a number of different approaches that try to recover some of this layer distributional context by computing an importance value for each connection that is dependent on the weights of the other connections in the same layer. We then generalise the LTH to use weight importances rather than weight magnitudes.

Experiments using these importance metrics on several architectures and datasets, reveal interesting aspects on the structure and emergence of Lottery tickets.

We find that given a repeatable training procedure, applying different importance metrics lead to distinct performant lottery tickets with little overlapping connections.
****
Generalized Sum Pooling for Metric Learning
Yeti Z. Gürbüz,Ozan Sener,A. Ayd■n Alatan
A common architectural choice for deep metric learning is a convolutional neural network followed by global average pooling (GAP). Albeit simple, GAP is a highly effective way to aggregate information. One possible explanation for the effectiveness of GAP is considering each feature vector as representing a different semantic entity and GAP as a convex combination of them. Following this perspective, we generalize GAP and propose a learnable generalized sum pooling method (GSP). GSP improves GAP with two distinct abilities: i) the ability to choose a subset of semantic entities, effectively learning to ignore nuisance information, and ii) learning the weights corresponding to the importance of each entity. Form

ally, we propose an entropy-smoothed optimal transport problem and show that it is a strict generalization of GAP, \ie a specific realization of the problem gives back GAP. We show that this optimization problem enjoys analytical gradients enabling us to use it as a direct learnable replacement for GAP. We further propose a zero-shot loss to ease the learning of GSP. We show the effectiveness of our method with extensive evaluations on 4 popular metric learning benchmarks. Code is available at: GSP-DML Framework

****

SAAL: Sharpness-Aware Active Learning

Yoon-Yeong Kim,JoonHo Jang,Byeonghu Na,Yeongmin Kim,Kyungwoo Song,Wanmo Kang,Il-chul Moon

While modern deep neural networks play significant roles in many research areas, they are also prone to overfitting problems under limited data instances. Particularly, this overfitting, or generalization issue, could be a problem in the framework of active learning because it selects a few data instances for learning over time. To consider the generalization, this paper introduces the first active learning method to incorporate the sharpness of loss space in the design of the acquisition function, inspired by sharpness-aware minimization (SAM). SAM intends to maximally perturb the training dataset, so the optimization can be led to a flat minima, which is known to have better generalization ability. Specifically, our active learning, Sharpness-Aware Active Learning (SAAL), constructs its acquisition function by selecting unlabeled instances whose perturbed loss becomes maximum. Over the adaptation of SAM into SAAL, we design a pseudo labeling mechanism to look forward to the perturbed loss w.r.t. the ground-truth label. Furthermore, we present a theoretic analysis between SAAL and recent active learning methods, so the recent works could be reduced to SAAL under a specific condition. We conduct experiments on various benchmark datasets for vision-based tasks in image classification and object detection. The experimental results confirm that SAAL outperforms the baselines by selecting instances that have the potentially maximal perturbation on the loss.

****

Scalable Subset Sampling with Neural Conditional Poisson Networks

Adeel Pervez,Phillip Lippe,Efstratios Gavves

A number of problems in learning can be formulated in terms of the basic primitive of sampling k elements out of a universe of n elements. This subset sampling operation cannot directly be included in differentiable models and approximations are essential. Current approaches take an \emph{order sampling} approach to sampling subsets and depend on differentiable approximations of the Top-k operator for selecting the largest k elements from a set. We present a simple alternative method for sampling subsets based on \emph{conditional Poisson sampling}. Unlike order sampling approaches, the parallel complexity of the proposed method is independent of the subset size which makes the method scalable to large subset sizes. We adapt the procedure to make it efficient and amenable to discrete gradient approximations for use in differentiable models. Furthermore, the method also allows the subset size parameter k to be differentiable. We demonstrate our approach on model explanation, image sub-sampling and stochastic k-nearest neighbor tasks outperforming existing methods in accuracy, efficiency and scalability.

****

Improved Convergence of Differential Private SGD with Gradient Clipping

Huang Fang,Xiaoyun Li,Chenglin Fan,Ping Li

Differential private stochastic gradient descent (DP-SGD) with gradient clipping (DP-SGD-GC) is an effective optimization algorithm that can train machine learning models with a privacy guarantee. Despite the popularity of DP-SGD-GC, its convergence in unbounded domain without the Lipschitz continuous assumption is less-understood; existing analysis of DP-SGD-GC either impose additional assumptions or end up with an utility bound that involves an non-vanishing bias term. In this work, for smooth and unconstrained problems, we improve the current analysis and show that DP-SGD-GC can achieve a vanishing utility bound without any bias term. Furthermore, when the noise generated from subsampled gradients is light-t ailed, we prove that DP-SGD-GC can achieve nearly the same utility bound as DP-SGD applies to the Lipschitz continuous objectives. As a by-product, we propose a new clipping technique, called value clipping, to mitigate the computational overhead caused by the classic gradient clipping. Experiments on standard benchmark datasets are conducted to support our analysis.

****

Group-level Brain Decoding with Deep Learning

Richard Csaky,Mats W.J. Van Es,Oiwi Parker Jones,Mark Woolrich

Decoding experimental variables from brain imaging data is gaining popularity, with applications in brain-computer interfaces and the study of neural representations. Decoding is typically subject-specific and does not generalise well over subjects. Here, we propose a method that uses subject embedding, analogous to word embedding in Natural Language Processing, to learn and exploit the structure in between subject variability as part of a decoding model, our adaptation of the WaveNet architecture for classification. We apply this to magnetoencephalography data, where 15 subjects viewed 118 different images, with 30 examples per image; to classify images using the entire 1s window following image presentation. We show that the combination of deep learning and subject embedding is crucial to closing the performance gap between subject- and group-level decoding models. Importantly, group models outperform subject models on low-accuracy subjects (but impair high-accuracy subjects) and can be helpful for initialising subject models. The potential of such group modelling is even higher with bigger datasets. To better enable physiological interpretation at the group level we demonstrate the use of permutation feature importance developing insights into the spatio-temporal and spectral information encoded in the models. All code is available on GitHub.

****

QUANTILE-LSTM: A ROBUST LSTM FOR ANOMALY DETECTION

Snehanshu Saha,Soma Dhavala,Jyotirmoy Sarkar,Preyank Bhavesh Mota,Santonu Sarkar

Anomalies refer to departure of systems and devices from their normal behaviour in standard operating conditions. An anomaly in an industrial device can indicate an upcoming failure, often in the temporal direction. In this paper, we make two contributions: 1) we estimate conditional quantiles, and consider three different ways to define anomalies based on the estimated quantiles and 2) use a new learnable activation function in the popular Long Short Term Memory (LSTM) architecture to model temporal long-range dependency. In particular, we propose Parametrized Elliot Function (Parametric Elliot Function (PEF)) as an activation function inside LSTM, which saturates lately compared to sigmoid and tanh. The proposed algorithms are compared with other well known anomaly detection algorithms, such as Isolation Forest (iForest), Elliptic Envelope, Autoencoder,and modern Deep Learning models such as Deep Autoencoding Gaussian Mixture Model (DAGMM), Generative Adversarial Networks (GAN) etc. The algorithms are evaluated in terms of various performance metrics, such as precision and recall. The algorithms are experimented on multiple industrial timeseries datasets such as Yahoo, AWS, GE, and machine sensor. We have found the LSTM based quantile algorithms are very effective and outperformed the existing algorithms in identifying the anomalies.

****

Neural Field Discovery Disentangles Equivariance in Interacting Dynamical Systems

Miltiadis Kofinas,Erik J Bekkers,Naveen Shankar Nagaraja,Efstratios Gavves

Systems of interacting objects often evolve under the influence of underlying field effects that govern their dynamics, \emph{e.g.} electromagnetic fields in physics, or map topologies and traffic rules in traffic scenes. While the interactions between objects depend on local information, the underlying fields depend on global states. Pedestrians and vehicles in traffic scenes, for example, follow different traffic rules and social norms depending on their absolute geolocation. The entanglement of global and local effects makes recently popularized equivariant networks inapplicable, since they fail to capture global information. To address this, in this work, we propose to \emph{disentangle} local object intera ctions --which are equivariant to global roto-translations and depend on relative positions and orientations-- from external global field effects --which depend on absolute positions and orientations. We theorize the presence of latent fields, which we aim to discover \emph{without} directly observing them, but infer them instead from the dynamics alone. We propose neural fields to learn the latent fields, and model the interactions with equivariant graph networks operating in local coordinate frames. We combine the two components in a graph network that transforms field effects in local frames and operates solely there. Our experiments show that we can accurately discover the underlying fields in charged particles settings, traffic scenes, and gravitational n-body problems, and effectively use them to learn the system and forecast future trajectories.

****

DIMENSION-REDUCED ADAPTIVE GRADIENT METHOD

Jingyang Li,Pan Zhou,Kuangyu Ding,Kim-Chuan Toh,Yinyu Ye

Adaptive gradient methods, such as Adam, have shown faster convergence speed than SGD across various kinds of network models. However, adaptive algorithms often suffer from inferior generalization performance than SGD. Though much effort via combining Adam and SGD have been invested to solve this issue, adaptive methods still fail to attain as good generalization as SGD. In this work, we proposed a Dimension-Reduced Adaptive Gradient Method (DRAG) to eliminate the generalization gap. DRAG makes an elegant combination of SGD and Adam by adopting a trust-region like framework. We observe that 1) Adam adjusts stepsizes for each gradient coordinate according to some loss curvature, and indeed decomposes the n-dimensional gradient into n independent directions to search, in which each direction inherits one coordinate element from the gradient and sets the remaining coordinate positions as zeros; 2) SGD uniformly scales gradient for all gradient coordinates and actually has only one descent direction to minimize. Accordingly, DRAG reduces the high degree of freedom of Adam and also improves the flexibility of SGD via optimizing the loss along $k\ (\ll \! n)$ descent directions, e.g. the gradient direction and momentum direction used in this work. Then per iteration, DRAG finds the best stepsizes for k descent directions by solving a trust-region subproblem whose computational overhead is negligible since the trust-region subproblem is low-dimensional, e.g. $k=2$ in this work. DRAG is compatible with the common deep learning training pipeline without introducing extra hyper-parameters and with negligible extra computation. Moreover, we prove the convergence property of DRAG for non-convex stochastic problems that often occur in deep learning training. Experimental results on representative benchmarks testify the fast convergence speed and also superior generalization of DRAG.

****

Learning to Estimate Single-View Volumetric Flow Motions without 3D Supervision

Erik Franz,Barbara Solenthaler,Nils Thuerey

We address the challenging problem of jointly inferring the 3D flow and volumetric densities moving in a fluid from a monocular input video with a deep neural network. Despite the complexity of this task, we show that it is possible to train the corresponding networks without requiring any 3D ground truth for training. In the absence of ground truth data we can train our model with observations from real-world capture setups instead of relying on synthetic reconstructions. We make this unsupervised training approach possible by first generating an initial prototype volume which is then moved and transported over time without the need for volumetric supervision. Our approach relies purely on image-based losses, an adversarial discriminator network, and regularization. Our method can estimate long-term sequences in a stable manner, while achieving closely matching targets for inputs such as rising smoke plumes.

****

ArCL: Enhancing Contrastive Learning with Augmentation-Robust Representations

Xuyang Zhao,Tianqi Du,Yisen Wang,Jun Yao,Weiran Huang

Self-Supervised Learning (SSL) is a paradigm that leverages unlabeled data for model training. Empirical studies show that SSL can achieve promising performance in distribution shift scenarios, where the downstream and training distributions differ. However, the theoretical understanding of its transferability remains limited. In this paper, we develop a theoretical framework to analyze the transferability of self-supervised contrastive learning, by investigating the impact of data augmentation on it. Our results reveal that the downstream performance of contrastive learning depends largely on the choice of data augmentation. Moreover, we show that contrastive learning fails to learn domain-invariant features, which limits its transferability. Based on these theoretical insights, we propose a novel method called Augmentation-robust Contrastive Learning (ArCL), which guarantees to learn domain-invariant features and can be easily integrated with existing contrastive learning algorithms. We conduct experiments on several datasets and show that ArCL significantly improves the transferability of contrastive learning.

****

Online Policy Optimization for Robust MDP

Jing Dong,Jingwei Li,Baoxiang Wang,Jingzhao Zhang

Reinforcement learning (RL) has exceeded human performance in many synthetic settings such as video games and Go. However, real-world deployment of end-to-end RL models is rare, as RL models can be very sensitive to slight perturbation of the environment. The robust Markov decision process (MDP) framework---in which the transition probabilities belong to an uncertainty set around a nominal model---provides one way to develop robust models. While previous analysis shows RL algorithms are effective assuming access to a generative model, it remains unclear whether RL can be efficient under a more realistic online setting, which requires carefully balancing exploration and exploitation. In this work, we consider online robust MDP by interacting with an unknown nominal system. We propose a robust optimistic policy optimization algorithm that is provably efficient. To address the additional uncertainty caused by an adversarial environment, our model features a new optimistic update rule derived via Fenchel conjugates. Our analysis establishes the first regret upper bound for online robust MDPs.

****

Toeplitz Neural Network for Sequence Modeling

Zhen Qin,Xiaodong Han,Weixuan Sun,Bowen He,Dong Li,Dongxu Li,Yuchao Dai,Lingpeng Kong,Yiran Zhong

Sequence modeling has important applications in natural language processing and computer vision. Recently, the transformer-based models have shown strong performance on various sequence modeling tasks, which rely on attention to capture pairwise token relations, and position embedding to inject positional information. While showing good performance, the transformer models are inefficient to scale to long input sequences, mainly due to the quadratic space-time complexity of attention. To overcome this inefficiency, we propose to model sequences with a relative position encoded Toeplitz matrix and use a Toeplitz matrix-vector production trick to reduce the space-time complexity of the sequence modeling to log linear. A lightweight sub-network called relative position encoder is proposed to generate relative position coefficients with a fixed budget of parameters, enabling the proposed Toeplitz neural network to deal with varying sequence lengths. In addition, despite being trained on 512-token sequences, our model can extrapolate input sequence length up to 14K tokens in inference with consistent performance. Extensive experiments on autoregressive and bidirectional language modeling, image modeling, and the challenging Long-range Arena Benchmark show that our method achieves better performance than its competitors in most downstream tasks while being significantly faster.

****

An Adaptive Entropy-Regularization Framework for Multi-Agent Reinforcement Learning

Woojun Kim,Youngchul Sung

In this paper, we propose an adaptive entropy-regularization framework (ADER) for multi-agent reinforcement learning (RL) to learn the adequate amount of exploration for each agent based on the degree of required exploration. In order to handle instability arising from updating multiple entropy temperature parameters for multiple agents, we disentangle the soft value function into two types: one for pure reward and the other for entropy. By applying multi-agent value factoriz ation to the disentangled value function of pure reward, we obtain a relevant metric to assess the necessary degree of exploration for each agent. Based on this metric, we propose the ADER algorithm based on maximum entropy RL, which controls the necessary level of exploration across agents over time by learning the proper target entropy for each agent. Experimental results show that the proposed scheme significantly outperforms current state-of-the-art multi-agent RL algorithms.

****

Relative Positional Encoding Family via Unitary Transformation

Zhen Qin,Weixuan Sun,Kaiyue Lu,Hui Deng,Dongxu Li,XiaoDong Han,Yuchao Dai,Lingpeng Kong,Yiran Zhong

Relative position encoding is widely used in vanilla and linear transformers to represent positional information. However, the existing encoding methods of a vanilla transformer are not always directly applicable to a linear transformer, because the latter requires a decomposition of the query and key representations into separate kernel functions. Nevertheless, principles to design encoding methods suitable for linear transformers remain under-studied. In this work, we put together a variety of existing encoding methods under a canonical form and further propose a family of relative positional encodings via unitary transformation. Our formulation leads to a principled framework that can be used to develop new relative positional encoding methods that preserve linear space-time complexity. Equipping with different parameters, the proposed unitary relative positional encoding family (URPE) derives effective encoding for various applications. Experiments show that compared with existing encoding methods, unitary encoding achieves competitive performance on language modeling and various challenging downstream tasks, such as machine translation and text classification. In the meantime, it highlights a general paradigm to design broadly more relative positional encoding methods, applicable inclusively to linear and vanilla transformers.

****

ColoristaNet for Photorealistic Video Style Transfer

Xiaowen Qiu,Ruize Xu,Boan He,Weifeng Ge

Photorealistic style transfer aims to transfer the artistic style of an image onto an input image or video while keeping photorealism. In this paper, we think it's the summary statistics matching scheme in existing algorithms that leads to unrealistic stylization. To avoid employing the popular Gram loss, we propose a self-supervised style transfer framework, which contains a style removal part and a style restoration part. The style removal network removes the original image styles, and the style restoration network recovers image styles in a supervised manner. Meanwhile, to address the problems in current feature transformation methods, we propose decoupled instance normalization to decompose feature transformation into style whitening and restylization. It works quite well in ColoristaNet and can transfer image styles efficiently while keeping photorealism. To ensure temporal coherency, we also incorporate optical flow methods and ConvLSTM to embed contextual information. Experiments demonstrates that ColoristaNet can achieve better stylization effects when compared with state-of-the-art algorithms.

****

Auto-Encoding Adversarial Imitation Learning

Kaifeng Zhang,Rui Zhao,Ziming Zhang,Yang Gao

Reinforcement learning (RL) provides a powerful framework for decision-making, but its application in practice often requires a carefully designed reward function. Adversarial Imitation Learning (AIL) sheds light on automatic policy acquisition without access to the reward signal from the environment. In this work, we propose Auto-Encoding Adversarial Imitation Learning (AEAIL), a robust and scalable AIL framework. To induce expert policies from demonstrations, AEAIL utilizes the reconstruction error of an auto-encoder as a reward signal, which provides more information for optimizing policies than the prior discriminator-based ones. Subsequently, we use the derived objective functions to train the auto-encoder and the agent policy. Experiments show that our AEAIL performs superior compared to state-of-the-art methods in the MuJoCo environments. More importantly, AEAIL shows much better robustness when the expert demonstrations are noisy. Specifi

cally, our method achieves 11% and 50.7% relative improvement overall compared to the best baseline GAIL and PWIL on clean and noisy expert data, respectively. Video results, open-source code and dataset are available in supplementary materials.
****

Δ-PINNs: physics-informed neural networks on complex geometries
Francisco Sahli Costabal,Simone Pezzuto,Paris Perdikaris
Physics-informed neural networks (PINNs) have demonstrated promise in solving forward and inverse problems involving partial differential equations. Despite recent progress on expanding the class of problems that can be tackled by PINNs, most of existing use-cases involve simple geometric domains. To date, there is no clear way to inform PINNs about the topology of the domain where the problem is being solved. In this work, we propose a novel positional encoding mechanism for PINNs based on the eigenfunctions of the Laplace-Beltrami operator. This technique allows to create an input space for the neural network that represents the geometry of a given object. We approximate the eigenfunctions as well as the operators involved in the partial differential equations with finite elements. We extensively test and compare the proposed methodology against traditional PINNs in complex shapes, such as a coil, a heat sink and a bunny, with different physics, such as the Eikonal equation and heat transfer. We also study the sensitivity of our method to the number of eigenfunctions used, as well as the discretization used for the eigenfunctions and the underlying operators. Our results show excellent agreement with the ground truth data in cases where traditional PINNs fail to produce a meaningful solution. We envision this new technique will expand the effectiveness of PINNs to more realistic applications.
****

BiTAT: Neural Network Binarization with Task-Dependent Aggregated Transformation
Geon Park,Jaehong Yoon,Haiyang Zhang,Xing Zhang,Sung Ju Hwang,Yonina C. Eldar
Neural network quantization aims to transform high-precision weights and activations of a given neural network into low-precision weights/activations for reduced memory usage and computation, while preserving the performance of the original model. However, extreme quantization (1-bit weight/1-bit activations) of compactly-designed backbone architectures (e.g., MobileNets) often used for edge-device deployments results in severe performance degeneration. This paper proposes a novel Quantization-Aware Training (QAT) method that can effectively alleviate performance degeneration even with extreme quantization by focusing on the inter-weight dependencies, between the weights within each layer and across consecutive layers. To minimize the quantization impact of each weight on others, we perform an orthonormal transformation of the weights at each layer by training an input-dependent correlation matrix and importance vector, such that each weight is disentangled from the others. Then, we quantize the weights based on their importance to minimize the loss of the information from the original weights/activations. We further perform progressive layer-wise quantization from the bottom layer to the top, so that quantization at each layer reflects the quantized distributions of weights and activations at previous layers. We validate the effectiveness of our method on various benchmark datasets against strong neural quantization baselines, demonstrating that it alleviates the performance degeneration on ImageNet and successfully preserves the full-precision model performance on CIFAR-100 with compact backbone networks.
****

Memory of Unimaginable Outcomes in Experience Replay
Adrian Remonda,Cole Corbitt Terrell,Eduardo E. Veas
Model-based reinforcement learning (MBRL) applies a single-shot dynamics model to imagined actions to select those with best expected outcome. The dynamics model is an unfaithful representation of the environment physics, and its capacity to predict the outcome of a future action varies as it is trained iteratively. An experience replay buffer collects the outcomes of all actions executed in the environment and is used to iteratively train the dynamics model. With growing experience, it is expected that the model becomes more accurate at predicting the outcome and expected reward of imagined actions. However, training times and memo

ry requirements drastically increase with the growing collection of experiences.

Indeed, it would be preferable to retain only those experiences that could not be anticipated by the model while interacting with the environment.
We argue that doing so results in a lean replay buffer with diverse experiences that correspond directly to the model's predictive weaknesses at a given point in time.
We propose strategies for: i) determining reliable predictions of the dynamics model with respect to the imagined actions, ii) retaining only the unimaginable experiences in the replay buffer, and iii) training further only when sufficient novel experience has been acquired.
We show that these contributions lead to lower training times, drastic reduction of the replay buffer size, fewer updates to the dynamics model and reduction of catastrophic forgetting. All of which enable the effective implementation of continual-learning agents using MBRL.
****

Temperature Schedules for self-supervised contrastive methods on long-tail data
Anna Kukleva,Moritz Böhle,Bernt Schiele,Hilde Kuehne,Christian Rupprecht
Most approaches for self-supervised learning (SSL) are optimised on curated balanced datasets, e.g. ImageNet, despite the fact that natural data usually exhibits long-tail distributions. In this paper, we analyse the behaviour of one of the most popular variants of SSL, i.e. contrastive methods, on imbalanced data. In particular, we investigate the role of the temperature parameter τ in the contrastive loss, by analysing the loss through the lens of average distance maximisation, and find that a large τ emphasises group-wise discrimination, whereas a small τ leads to a higher degree of instance discrimination. While τ has thus far been treated exclusively as a constant hyperparameter, in this work, we propose to employ a dynamic τ and show that a simple cosine schedule can yield significant improvements in the learnt representations. Such a schedule results in a constant `task switching' between an emphasis on instance discrimination and group-wise discrimination and thereby ensures that the model learns both group-wise features, as well as instance-specific details. Since frequent classes benefit from the former, while infrequent classes require the latter, we find this method to consistently improve separation between the classes in long-tail data without any additional computational cost.
****

Deep Learning on Implicit Neural Representations of Shapes
Luca De Luigi,Adriano Cardace,Riccardo Spezialetti,Pierluigi Zama Ramirez,Samuele Salti,Luigi di Stefano
Implicit Neural Representations (INRs) have emerged in the last few years as a powerful tool to encode continuously a variety of different signals like images, videos, audio and 3D shapes. When applied to 3D shapes, INRs allow to overcome the fragmentation and shortcomings of the popular discrete representations used so far. Yet, considering that INRs consist in neural networks, it is not clear whether and how it may be possible to feed them into deep learning pipelines aimed at solving a downstream task. In this paper, we put forward this research problem and propose inr2vec, a framework that can compute a compact latent representation for an input INR in a single inference pass. We verify that inr2vec can embed effectively the 3D shapes represented by the input INRs and show how the produced embeddings can be fed into deep learning pipelines to solve several tasks by processing exclusively INRs.
****

Continual Vision-Language Representaion Learning with Off-Diagonal Information
Zixuan Ni,Longhui Wei,Siliang Tang,Yueting Zhuang,Qi Tian
Multimodal pre-trained methods with a contrastive learning framework (like CLIP) have recently achieved consistent advantages on various cross-model downstream tasks. However, they usually require a large amount of image-text samples and a vast computing budget for training, which makes the re-training process expensive while the training data is collected continuously (the phenomenon is widespread in real scenarios). In this paper, we discuss the feasibility of continuously

training CLIP models based on discrete streaming data. We find that the multimodal retrieval performance of the CLIP in a continual training setting is significantly lower than that in a joint training setting. We name this phenomenon Cognitive Disorder(CD). By tracking the directional changes of the representation vectors in the continuously updated CLIP model, we explore and summarize the spatial variation of the modal encoders within the CLIP: Intra-modal Rotation and Inter-modal Deviation. Intra-modal Rotation means that the vision and language representation space in the CLIP is rotating greatly around the center of a high-dimensional unit sphere during continual training, accompanied by a relatively small change in the topology of the representation space. Inter-modal deviation happens when the vision and language's intra-modal rotation is unsynchronized. Moreover, we empirically and theoretically demonstrate how intra-modal rotation and inter-modal deviation lead to CD. In order to alleviate CD in continual CLIP training, we propose a new continual training framework Mod-X: Maintain off-diagonal information-matrix. By selectively aligning the off-diagonal information distribution of contrastive matrixes, the Mod-X helps the model not only better fits the newly trained data domain but also maintains the multimodal cognitive ability on the old data domain during the continual large-scale training (Section \ref{experiments}).

****

Learning Counterfactually Invariant Predictors

Francesco Quinzan,Cecilia Casolo,Krikamol Muandet,Niki Kilbertus,Yucen Luo

We propose a method to learn predictors that are invariant under counterfactual changes of certain covariates. This method is useful when the prediction target is causally influenced by covariates that should not affect the predictor output. For instance, this could prevent an object recognition model from being influenced by position, orientation, or scale of the object itself. We propose a model-agnostic regularization term based on conditional kernel mean embeddings to enforce counterfactual invariance during training. We prove the soundness of our method, which can handle mixed categorical and continuous multivariate attributes. Empirical results on synthetic and real-world data demonstrate the efficacy of our method in a variety of settings.

****

ImaginaryNet: Learning Object Detectors without Real Images and Annotations

Minheng Ni,Zitong Huang,Kailai Feng,Wangmeng Zuo

Without the demand of training in reality, humans are able of detecting a new category of object simply based on the language description on its visual characteristics. Empowering deep learning with this ability undoubtedly enables the neural network to handle complex vision tasks, e.g., object detection, without collecting and annotating real images. To this end, this paper introduces a novel challenging learning paradigm Imaginary-Supervised Object Detection (ISOD), where neither real images nor manual annotations are allowed for training object detectors. To resolve this challenge, we propose ImaginaryNet, a framework to synthesize images by combining pretrained language model and text-to-image synthesis model. Given a class label, the language model is used to generate a full description of a scene with a target object, and the text-to-image model is deployed to generate a photo-realistic image. With the synthesized images and class labels, weakly supervised object detection can then be leveraged to accomplish ISOD. By gradually introducing real images and manual annotations, ImaginaryNet can collaborate with other supervision settings to further boost detection performance. Experiments show that ImaginaryNet can (i) obtain about 75% performance in ISOD compared with the weakly supervised counterpart of the same backbone trained on real data, (ii) significantly improve the baseline while achieving state-of-the-art or comparable performance by incorporating ImaginaryNet with other supervision settings. Our code will be publicly available at https://github.com/kodenii/ImaginaryNet.

****

Don't Throw Your Old Policies Away: Knowledge-based Policy Recycling Protects Against Adversarial Attacks

Yaqi Xie,Chen Yu,Harold Soh

Recent work has shown that Deep Reinforcement Learning (DRL) is vulnerable to adversarial attacks, in which minor perturbations of input signals cause agents to behave inappropriately and unexpectedly. Humans, on the other hand, appear robust to these particular sorts of input variations. We posit that this part of robustness stems from accumulated knowledge about the world.
In this work, we propose to leverage prior knowledge to defend against adversarial attacks in RL settings using a framework we call Knowledge-based Policy Recycling (KPR). Different from previous defense methods such as adversarial training and robust learning, KPR incorporates domain knowledge over a set of auxiliary tasks policies and learns relations among them from interactions with the environment via a Graph Neural Network (GNN). KPR can use any relevant policy as an auxiliary policy and, importantly, does not assume access or information regarding the adversarial attack. Empirically, KPR results in policies that are more robust to various adversarial attacks in Atari games and a simulated Robot Foodcourt environment.
****

Contextual bandits with concave rewards, and an application to fair ranking
Virginie Do,Elvis Dohmatob,Matteo Pirotta,Alessandro Lazaric,Nicolas Usunier
We consider Contextual Bandits with Concave Rewards (CBCR), a multi-objective bandit problem where the desired trade-off between the rewards is defined by a known concave objective function, and the reward vector depends on an observed stochastic context. We present the first algorithm with provably vanishing regret for CBCR without restrictions on the policy space, whereas prior works were restricted to finite policy spaces or tabular representations. Our solution is based on a geometric interpretation of CBCR algorithms as optimization algorithms over the convex set of expected rewards spanned by all stochastic policies. Building on Frank-Wolfe analyses in constrained convex optimization, we derive a novel reduction from the CBCR regret to the regret of a \emph{scalar-reward} bandit problem. We illustrate how to apply the reduction off-the-shelf to obtain algorithms for CBCR with both linear and general reward functions, in the case of non-combinatorial actions. Motivated by fairness in recommendation, we describe a special case of CBCR with rankings and fairness-aware objectives, leading to the first algorithm with regret guarantees for contextual combinatorial bandits with fairness of exposure.
****

Gradient Boosting Performs Gaussian Process Inference
Aleksei Ustimenko,Artem Beliakov,Liudmila Prokhorenkova
This paper shows that gradient boosting based on symmetric decision trees can be equivalently reformulated as a kernel method that converges to the solution of a certain Kernel Ridge Regression problem. Thus, we obtain the convergence to a Gaussian Process' posterior mean, which, in turn, allows us to easily transform gradient boosting into a sampler from the posterior to provide better knowledge uncertainty estimates through Monte-Carlo estimation of the posterior variance. We show that the proposed sampler allows for better knowledge uncertainty estimates leading to improved out-of-domain detection.
****

Constrained Reinforcement Learning for Safety-Critical Tasks via Scenario-Based Programming
Davide Corsi,Raz Yerushalmi,Guy Amir,Alessandro Farinelli,David Harel,Guy Katz
Deep reinforcement learning (DRL) has achieved groundbreaking successes in various applications, including robotics. A natural consequence is the adoption of this paradigm for safety-critical tasks, where human safety and expensive hardware can be involved. In this context, it is crucial to optimize the performance of DRL-based agents while providing guarantees about their behavior. This paper presents a novel technique for incorporating domain-expert knowledge into a constrained DRL training loop. Our technique exploits the scenario-based programming paradigm, designed to specify such knowledge in a simple and intuitive way. While our approach can be considered general purpose, we validated our method by performing experiments on a synthetic set of benchmark environments, and the popular robotic mapless navigation problem, in simulation and on the actual platform. Ou

r results demonstrate that using our approach to leverage expert knowledge dramatically improves the safety and performance of the agent.
****

When is Adversarial Robustness Transferable?

Anna-Kathrin Kopetzki,Aleksandar Bojchevski,Stephan Günnemann

Knowledge transfer is an effective tool for learning, especially when labeled data is scarce or when training from scratch is prohibitively costly. The overwhelming majority of transfer learning literature is focused on obtaining accurate models, neglecting the issue of adversarial robustness. Yet, robustness is essential, particularly when transferring to safety-critical domains.

We analyze and compare how different training procedures on the source domain and different fine-tuning strategies on the target domain affect robustness. More precisely, we study 10 training schemes for source models and 3 for target models, including normal, adversarial, contrastive and Lipschitz constrained variants. We quantify model robustness via randomized smoothing and adversarial attacks. Our results show that improving model robustness on the source domain increases robustness on the target domain. Target retraining has a minor influence on target model robustness. These results indicate that model robustness is preserved during target retraining and transfered from the source domain to the target domain.
****

COFS: COntrollable Furniture layout Synthesis

Wamiq Reyaz Para,Paul Guerrero,Niloy Mitra,Peter Wonka

Realistic, scalable, and controllable generation of furniture layouts is essential for many applications in virtual reality, augmented reality, game development and synthetic data generation. The most successful current methods tackle this problem as a sequence generation problem which imposes a specific ordering on the elements of the layout, making it hard to exert fine-grained control over the attributes of a generated scene. Existing methods provide control through object-level conditioning, or scene completion, where generation can be conditioned on an arbitrary subset of furniture objects. However, attribute-level conditioning, where generation can be conditioned on an arbitrary subset of object attributes, is not supported. We propose COFS, a method to generate furniture layouts that enables fine-grained control through attribute-level conditioning. For example, COFS allows specifying only the scale and type of objects that should be placed in the scene and the generator chooses their positions and orientations; or the position that should be occupied by objects can be specified and the generator chooses their type, scale, orientation, etc. Our results show both qualitatively and quantitatively that we significantly outperform existing methods on attribute-level conditioning.
****

Distribution Shift Detection for Deep Neural Networks

Guy Bar-Shalom,Yonatan Geifman,Ran El-Yaniv

To deploy and operate deep neural models in production, the quality of their predictions, which might be contaminated benignly or manipulated maliciously by input distributional deviations, must be monitored and assessed. Specifically, we study the case of monitoring the healthy operation of a deep neural network (DNN) receiving a stream of data, with the aim of detecting input distributional deviations over which the quality of the network's predictions is potentially damaged. Using selective prediction principles, we propose a distribution deviation detection method for DNNs. The proposed method is derived from a tight coverage generalization bound computed over a sample of instances drawn from the true underlying distribution. Based on this bound, our detector continuously monitors the operation of the network over a test window and fires off an alarm whenever a deviation is detected. This novel detection method consistently and significantly outperforms the state of the art with respect to the CIFAR-10 and ImageNet datasets, thus establishing a new performance bar for this task, while being substantially more efficient in time and space complexities.
****

Learning Zero-Shot Cooperation with Humans, Assuming Humans Are Biased

Chao Yu,Jiaxuan Gao,Weilin Liu,Botian Xu,Hao Tang,Jiaqi Yang,Yu Wang,Yi Wu

There is a recent trend of applying multi-agent reinforcement learning (MARL) to train an agent that can cooperate with humans in a zero-shot fashion without using any human data. The typical workflow is to first repeatedly run self-play (SP) to build a policy pool and then train the final adaptive policy against this pool. A crucial limitation of this framework is that every policy in the pool is optimized w.r.t. the environment reward function, which implicitly assumes that the testing partners of the adaptive policy will be precisely optimizing the same reward function as well. However, human objectives are often substantially biased according to their own preferences, which can differ greatly from the environment reward. We propose a more general framework, Hidden-Utility Self-Play (HSP), which explicitly models human biases as hidden reward functions in the self-play objective. By approximating the reward space as linear functions, HSP adopts an effective technique to generate an augmented policy pool with biased policies. We evaluate HSP on the Overcooked benchmark. Empirical results show that our HSP method produces higher rewards than baselines when cooperating with learned human models, manually scripted policies, and real humans. The HSP policy is also rated as the most assistive policy based on human feedback.

****

SUG: Single-dataset Unified Generalization for 3D Point Cloud Classification

Siyuan Huang,Bo Zhang,Botian Shi,Peng Gao,Tao Chen,Hongsheng Li,Yikang LI

In recent years, research on zero-shot domain adaptation, namely Domain Generalization (DG), which aims to adapt a well-trained source domain model to unseen target domains without accessing any target sample, has been fast-growing in the 2D image tasks such as classification and object detection. However, its exploration on 3D point cloud data is still insufficient and challenged by more complex and uncertain cross-domain variances with irregular point data structures and uneven inter-class modality distribution. In this paper, different from previous 2D DG works, we focus on the 3D DG problem, and propose a Single-dataset Unified Generalization (SUG) framework that only leverages the source domain data to alleviate the unforeseen domain differences faced by the well-pretrained source model. Specifically, we first design a Multi-grained Sub-domain Alignment (MSA) method that can constrain the learned representations to be domain-agnostic and discriminative, by performing a multi-grained feature alignment process between the splitted sub-domains from the single source dataset. Then, a Sample-level Domain-aware Attention (SDA) strategy is presented, which can selectively enhance easy-to-adapt samples from different sub-domains according to the sample-level inter-domain distance, to avoid the negative transfer. Extensive experiments are conducted on three common 3D point cloud benchmarks. The experimental results demonstrate that SUG framework is effective to boost the model generalization ability for unseen target domains, even outperforming the existing unsupervised domain adaptation methods that have to access extensive target domain data, where we significantly improve classification accuracy by 7.7% on ModelNet-to-ScanNet setting and 2.3% on ShapeNet-to-ScanNet setting. Our code will be available.

****

An Optimal Transport Perspective on Unpaired Image Super-Resolution

Milena Gazdieva,Litu Rout,Alexander Korotin,Andrey Kravchenko,Alexander Filippov,Evgeny Burnaev

Real-world image super-resolution (SR) tasks often do not have paired
datasets, which limits the application of supervised techniques. As a result, the tasks are usually approached by unpaired techniques based on Generative Adversarial Networks (GANs), which yield complex training losses with several regularization terms, e.g., content or identity losses. We theoretically investigate optimization problems which arise in such models and find two surprizing observations. First, the learned SR map is always an optimal transport (OT) map. Second, we theoretically prove and empirically show that the learned map is biased, i.e., it does not actually transform the distribution of low-resolution images to high-resolution ones. Inspired by these findings, we propose an algorithm for unpaired SR which learns an unbiased OT map for the perceptual transport cost. Unlike the existing GAN-based alternatives, our algorithm has a simple optimization ob

jective reducing the need for complex hyperparameter selection and an application of additional regularizations. At the same time, it provides a nearly state-of-the-art performance on the large-scale unpaired AIM19 dataset.
****

A Functional Perspective on Multi-Layer Out-of-Distribution Detection
Eduardo Dadalto Câmara Gomes,Pierre Colombo,Guillaume Staerman,Nathan Noiry,Pablo Piantanida
A crucial component for implementing reliable classifiers is detecting examples far from the reference (training) distribution, referred to as out-of-distribution (OOD) samples. A key feature of OOD detection is to exploit the network by extracting statistical patterns and relationships through the pre-trained multi-layer classifier. Despite achieving solid results, state-of-the-art methods require either additional OOD examples, expensive computation of gradients, or are tightened to a particular architecture, limiting their applications. This work adopts an original approach based on a functional view of the network that exploits the sample's trajectories through the various layers and their statistical dependencies. In this new framework, OOD detection translates into detecting samples whose trajectories differ from the typical behavior characterized by the training set. Our method significantly decreases the OOD detection error of classifiers trained on ImageNet and outperforms the state-of-the-art methods on average AUROC and TNR at 95% TPR. We demonstrate that the functional signature left by a sample in a network carries relevant information for OOD detection.
****

Modelling Long Range Dependencies in ND: From Task-Specific to a General Purpose CNN
David M Knigge,David W. Romero,Albert Gu,Efstratios Gavves,Erik J Bekkers,Jakub Mikolaj Tomczak,Mark Hoogendoorn,Jan-jakob Sonke
Performant Convolutional Neural Network (CNN) architectures must be tailored to specific tasks in order to consider the length, resolution, and dimensionality of the input data. In this work, we tackle the need for problem-specific CNN architectures. We present the Continuous Convolutional Neural Network (CCNN): a single CNN able to process data of arbitrary resolution, dimensionality and length without any structural changes. Its key component are its continuous convolutional kernels which model long-range dependencies at every layer, and thus remove the need of current CNN architectures for task-dependent downsampling and depths. We showcase the generality of our method by using the same architecture for tasks on sequential ($1{\rm D}$), visual ($2{\rm D}$) and point-cloud ($3{\rm D}$) data. Our CCNN matches and often outperforms the current state-of-the-art across all tasks considered.
****

Ahead-of-Time P-Tuning
Daniil Gavrilov
This paper proposes a new parameter-efficient method for fine-tuning, AoT P-Tuning. This method adds input-dependent biases before evaluating the Transformer layer, reducing the required evaluation time when compared to P-Tuning. Same as P-Tuning, AoT P-Tuning allows multi-task inference with a single backbone model for evaluating different tasks in a single batch.
We experimented with the proposed method on the GLUE and SuperGLUE benchmarking datasets using RoBERTa-Base, RoBERTa-Large, and DeBERTa-XL backbone models. Our observations show that AoT P-tuning performed on par with or better than P-Tuning v2 while being up to $1.3\times$ times faster during inference.
****

Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training
Simone Zini,Alex Gomez-Villa,Marco Buzzelli,Bart■omiej Twardowski,Andrew D. Bagdanov,Joost van de weijer
Several recent works on self-supervised learning are trained by mapping different augmentations of the same image to the same feature representation. The data augmentations used are of crucial importance to the quality of learned feature representations. In this paper, we analyze how the color jitter traditionally used

in data augmentation negatively impacts the quality of the color features in learned feature representations. To address this problem, we propose a more realistic, physics-based color data augmentation - which we call Planckian Jitter - that creates realistic variations in chromaticity and produces a model robust to illumination changes that can be commonly observed in real life, while maintaining the ability to discriminate image content based on color information.

Experiments confirm that such a representation is complementary to the representations learned with the currently-used color jitter augmentation and that a simple concatenation leads to significant performance gains on a wide range of downstream datasets.

In addition, we present a color sensitivity analysis that documents the impact of different training methods on model neurons and shows that the performance of the learned features is robust with respect to illuminant variations.

Official code available at: https://github.com/TheZino/PlanckianJitter

****

Efficient and Stealthy Backdoor Attack Triggers are Close at Hand

Minlong Peng,Xu Li,Mingming Sun

A backdoor attack aims to inject a backdoor into a deep model so that the model performs normally on benign samples while maliciously predicting the input as the attacker-defined target class when the backdoor is activated by a predefined trigger pattern. Most existing backdoor attacks use a pattern that rarely occurs in benign data as the trigger pattern. In this way, the impact of the attack on the label prediction of benign data can be mitigated. However, this practice also results in the attack being defended against with little performance degradation on benign data by preventing the trigger pattern from being activated. In this work, we present a new attack strategy to solve this dilemma. Unlike the conventional strategy, our strategy extracts the trigger pattern from benign training data, which frequently occurs in samples of the target class but rarely occurs in samples of the other classes. Compared with the prevailing strategy, our proposed strategy has two advantages. First, it can improve the efficiency of the attack because learning on benign samples of the target class can facilitate the fitting of the trigger pattern. Second, it increases the difficulty or cost of identifying the trigger pattern and preventing its activation, since many benign samples of the target class contain the trigger pattern. We empirically evaluate our strategy on four benchmark datasets. The experimental studies show that attacks performed with our strategy can achieve much better performance when poisoning only 0.1\% or more of the training data, and can achieve better performance against several benchmark defense algorithms.

****

Property Inference Attacks Against t-SNE Plots

Jingcan Chen,Xinlei He,Boyang Zhang,Yang Chen,Yang Zhang

With the prevailing of machine learning (ML), researchers have shown that ML models are also vulnerable to various privacy and security attacks. As one of the representative attacks, the property inference attack aims to infer the private/sensitive properties of the training data (e.g., race distribution) given the output of ML models. In this paper, we present a new side channel for property inference attacks, i.e., t-SNE plots, which are widely used to show feature distribution or demonstrate model performance. We show for the first time that the private/sensitive properties of the data that are used to generate the plot can be successfully predicted. Briefly, we leverage the publicly available model as the shadow model to generate t-SNE plots with different properties. We use those plots to train an attack model, which is a simple image classifier, to infer the specific property of a given t-SNE plot. Extensive evaluation on four datasets shows that our proposed attack can effectively infer the undisclosed property of the data presented in the t-SNE plots, even when the shadow model is different from the target model used to generate the t-SNE plots. We also reveal that the attacks are robust in various scenarios, such as constructing the attack with fewer t-SNE plots/different density settings and attacking t-SNE plots generated by fine-tuned target models. The simplicity of our attack method indicates that the potential risk of leaking sensitive properties in t-SNE plots is largely underest

imated. As possible defenses, we observe that adding noise to the image embeddings or t-SNE coordinates effectively mitigates attacks but can be bypassed by adaptive attacks, which prompts the need for more effective defenses.
****

GAMR: A Guided Attention Model for (visual) Reasoning
Mohit Vaishnav,Thomas Serre

Humans continue to outperform modern AI systems in their ability to flexibly parse and understand complex visual scenes. Here, we present a novel module for visual reasoning, the Guided Attention Model for (visual) Reasoning ($\textit{GAMR}$), which instantiates an active vision theory -- positing that the brain solves complex visual reasoning problems dynamically -- via sequences of attention shifts to select and route task-relevant visual information into memory. Experiments on an array of visual reasoning tasks and datasets demonstrate GAMR's ability to learn visual routines in a robust and sample-efficient manner. In addition, GAMR is shown to be capable of zero-shot generalization on completely novel reasoning tasks. Overall, our work provides computational support for cognitive theories that postulate the need for a critical interplay between attention and memory to dynamically maintain and manipulate task-relevant visual information to solve complex visual reasoning tasks.
****

Voint Cloud: Multi-View Point Cloud Representation for 3D Understanding
Abdullah Hamdi,Silvio Giancola,Bernard Ghanem

Multi-view projection methods have demonstrated promising performance on 3D understanding tasks like 3D classification and segmentation. However, it remains unclear how to combine such multi-view methods with the widely available 3D point clouds. Previous methods use unlearned heuristics to combine features at the point level. To this end, we introduce the concept of the multi-view point cloud (Voint cloud), representing each 3D point as a set of features extracted from several view-points. This novel 3D Voint cloud representation combines the compactness of 3D point cloud representation with the natural view-awareness of multi-view representation. Naturally, we can equip this new representation with convolutional and pooling operations. We deploy a Voint neural network (VointNet) to learn representations in the Voint space. Our novel representation achieves state-of-the-art performance on 3D classification, shape retrieval, and robust 3D part segmentation on standard benchmarks (ScanObjectNN, ShapeNet Core55, and ShapeNet Parts). Further analysis shows that VointNet improves the robustness to occlusion compared to other methods.
****

QuAnt: Quantum Annealing with Learnt Couplings
Marcel Seelbach Benkner,Maximilian Krahn,Edith Tretschk,Zorah Lähner,Michael Moeller,Vladislav Golyanik

Modern quantum annealers can find high-quality solutions to combinatorial optimisation objectives given as quadratic unconstrained binary optimisation (QUBO) problems. Unfortunately, obtaining suitable QUBO forms in computer vision remains challenging and currently requires problem-specific analytical derivations. Moreover, such explicit formulations impose tangible constraints on solution encodings. In stark contrast to prior work, this paper proposes to learn QUBO forms from data through gradient backpropagation instead of deriving them. As a result, the solution encodings can be chosen flexibly and compactly. Furthermore, our methodology is general and virtually independent of the specifics of the target problem type. We demonstrate the advantages of learnt QUBOs on the diverse problem types of graph matching, 2D point cloud alignment and 3D rotation estimation. Our results are competitive with the previous quantum state of the art while requiring much fewer logical and physical qubits, enabling our method to scale to larger problems. The code and the new dataset are available at https://4dqv.mpi-inf.mpg.de/QuAnt/.
****

Understanding Gradient Regularization in Deep Learning: Efficient Finite-Difference Computation and Implicit Bias
Ryo Karakida,Tomoumi Takase,Tomohiro Hayase,Kazuki Osawa

Gradient regularization (GR) is a method that penalizes the gradient norm of the training loss during training. Although some studies have reported that GR improves generalization performance in deep learning, little attention has been paid to it from the algorithmic perspective, that is, the algorithms of GR that efficiently improve performance. In this study, we first reveal that a specific finite-difference computation, composed of both gradient ascent and descent steps, reduces the computational cost for GR. In addition, this computation empirically achieves better generalization performance. Next, we theoretically analyze a solvable model, a diagonal linear network, and clarify that GR has a desirable implicit bias. Learning with GR chooses better minima in a certain problem, and the finite-difference GR chooses even better ones as the ascent step size becomes larger. Finally, we demonstrate that finite-difference GR is closely related to some other algorithms based on iterative ascent and descent steps for exploring flat minima: sharpness-aware minimization and the flooding method. In particular, we reveal that flooding performs finite-difference GR in an implicit way. Thus, this work broadens our understanding of GR in both practice and theory.
****
Approximate Nearest Neighbor Search through Modern Error-Correcting Codes
Noam Touitou,Nissim Halabi
A locality-sensitive hash (or LSH) is a function that can efficiently map dataset points into a latent space while preserving pairwise distances. Such LSH functions have been used in approximate nearest-neighbor search (ANNS) in the following classic way, which we call classic hash clustering (CHC): first, the dataset points are hashed into a low-dimensional binary space using the LSH function; then, the points are clustered by these hash values. Upon receiving a query, its nearest neighbors are sought within its hash-cluster and nearby hash-clusters (i.e., multi-probe). However, CHC mandates a low-dimensional latent space for the LSH function, which distorts distances from the (high-dimensional) original real space; this results in inferior recall. This is often mitigated through using multiple hash tables at additional storage and memory costs.

In this paper, we introduce a better way of using LSH functions for ANNS. Our method, called the Polar Code Nearest-Neighbor (PCNN) algorithm, uses modern error-correcting codes (specifically polar codes) to maintain a manageable number of clusters inside a high-dimensional latent space. Allowing the LSH function to embed into this high-dimensional latent space results in higher recall, as the embedding faithfully captures distances in the original space. The crux of PCNN is using polar codes for probing: we present a multi-probe scheme for PCNN which uses efficient list-decoding methods for polar codes, with time complexity independent of the dataset size. Fixing the choice of LSH, experiment results demonstrate significant performance gains of PCNN over CHC; in particular, PCNN with a single table outperforms CHC with multiple tables, obviating the need for large memory and storage.
****
Social and environmental impact of recent developments in machine learning on biology and chemistry research
Daniel Probst
Potential societal and environmental effects such as the rapidly increasing resource use and the associated environmental impact, reproducibility issues, and exclusivity, the privatization of ML research leading to a public research brain-drain, a narrowing of the research effort caused by a focus on deep learning, and the introduction of biases through a lack of sociodemographic diversity in data and personnel caused by recent developments in machine learning are a current topic of discussion and scientific publications. However, these discussions and publications focus mainly on computer science-adjacent fields, including computer vision and natural language processing or basic ML research. Using bibliometric analysis of the complete and full-text analysis of the open-access literature, we show that the same observations can be made for applied machine learning in chemistry and biology. These developments can potentially affect basic and applied research, such as drug discovery and development, beyond the known issue of bi

ased data sets.
****

When to Make and Break Commitments?

Alihan Hüyük,Zhaozhi Qian,Mihaela van der Schaar

In many scenarios, decision-makers must commit to long-term actions until their resolution before receiving the payoff of said actions, and usually, staying committed to such actions incurs continual costs. For instance, in healthcare, a newly-discovered treatment cannot be marketed to patients until a clinical trial is conducted, which both requires time and is also costly. Of course in such scenarios, not all commitments eventually pay off. For instance, a clinical trial might end up failing to show efficacy. Given the time pressure created by the continual cost of keeping a commitment, we aim to answer: When should a decision-maker break a commitment that is likely to fail—either to make an alternative commitment or to make no further commitments at all? First, we formulate this question as a new type of optimal stopping/switching problem called the optimal commitment problem (OCP). Then, we theoretically analyze OCP, and based on the insights we gain, propose a practical algorithm for solving it. Finally, we empirically evaluate the performance of our algorithm in running clinical trials with subpopulation selection.

****

Generalization bounds and algorithms for estimating the effect of multiple treatments and dosage

Alexis Bellot,Anish Dhir,Giulia Prando

Estimating conditional treatment effects has been a longstanding challenge for fields of study such as epidemiology or economics that require a treatment-dosage pair to make decisions, but may not be able to run randomized trials to precisely quantify their effect. In the context of representation learning, there is an extensive literature relating model architectures with regularization techniques to solve this problem using observational data. However, theoretically motivated loss functions and bounds on generalization errors only exist in select circumstances, such as in the presence of binary treatments. In this paper, we introduce new bounds on the counterfactual generalization error in the context of multiple treatments and continuous dosage parameters, which subsume existing results. This result, in a principled manner, guides the definition of new learning objectives that can be used to train representation learning algorithms. We show empirically new state-of-the-art performance results across several benchmark data sets for this problem, including in comparison to doubly-robust estimation methods.

****

DENSE RGB SLAM WITH NEURAL IMPLICIT MAPS

Heng Li,Xiaodong Gu,Weihao Yuan,luwei yang,Zilong Dong,Ping Tan

There is an emerging trend of using neural implicit functions for map representation in Simultaneous Localization and Mapping (SLAM). Some pioneer works have achieved encouraging results on RGB-D SLAM. In this paper, we present a dense RGB SLAM method with neural implicit map representation. To reach this challenging goal without depth input, we introduce a hierarchical feature volume to facilitate the implicit map decoder. This design effectively fuses shape cues across different scales to facilitate map reconstruction. Our method simultaneously solves the camera motion and the neural implicit map by matching the rendered and input video frames. To facilitate optimization, we further propose a photometric warping loss in the spirit of multi-view stereo to better constrain the camera pose and scene geometry. We evaluate our method on commonly used benchmarks and compare it with modern RGB and RGB-D SLAM systems. Our method achieves favorable results than previous methods and even surpasses some recent RGB-D SLAM methods.The code is at poptree.github.io/DIM-SLAM/.

****

Monocular Scene Reconstruction with 3D SDF Transformers

Weihao Yuan,Xiaodong Gu,Heng Li,Zilong Dong,Siyu Zhu

Monocular scene reconstruction from posed images is challenging due to the complexity of a large environment. Recent volumetric methods learn to directly predic t the TSDF volume and have demonstrated promising results in this task. However, most methods focus on how to extract and fuse the 2D features to a 3D feature volume, but none of them improve the way how the 3D volume is aggregated. In this work, we propose an SDF transformer network, which replaces the role of 3D CNN for better 3D feature aggregation. To reduce the explosive computation complexity of the 3D multi-head attention, we propose a sparse window attention module, where the attention is only calculated between the non-empty voxels within a local window. Then a top-down-bottom-up 3D attention network is built for 3D feature aggregation, where a dilate-attention structure is proposed to prevent geometry degeneration, and two global modules are employed to equip with global receptive fields. The experiments on multiple datasets show that this 3D transformer network generates a more accurate and complete reconstruction, which outperforms previous methods by a large margin. Remarkably, the mesh accuracy is improved by 41.8%, and the mesh completeness is improved by 25.3% on the ScanNet dataset. The code of our method will be made public.

****

Learning Heterogeneous Interaction Strengths by Trajectory Prediction with Graph Neural Network

Seungwoong Ha,Hawoong Jeong

Dynamical systems with interacting agents are universal in nature, commonly modeled by a graph of relationships between their constituents. Recently, various works have been presented to tackle the problem of inferring those relationships from the system trajectories via deep neural networks, but most of the studies assume binary or discrete types of interactions for simplicity. In the real world, the interaction kernels often involve continuous interaction strengths, which cannot be accurately approximated by discrete relations. In this work, we propose the relational attentive inference network (RAIN) to infer continuously weighted interaction graphs without any ground-truth interaction strengths. Our model employs a novel pairwise attention (PA) mechanism to refine the trajectory representations and a graph transformer to extract heterogeneous interaction weights for each pair of agents. We show that our RAIN model with the PA mechanism accurately infers continuous interaction strengths for simulated physical systems in an unsupervised manner. Further, RAIN with PA successfully predicts trajectories from motion capture data with an interpretable interaction graph, demonstrating the virtue of modeling unknown dynamics with continuous weights.

****

From t-SNE to UMAP with contrastive learning

Sebastian Damrich,Niklas Böhm,Fred A Hamprecht,Dmitry Kobak

Neighbor embedding methods t-SNE and UMAP are the de facto standard for visualizing high-dimensional datasets. Motivated from entirely different viewpoints, their loss functions appear to be unrelated. In practice, they yield strongly differing embeddings and can suggest conflicting interpretations of the same data. The fundamental reasons for this and, more generally, the exact relationship between t-SNE and UMAP have remained unclear. In this work, we uncover their conceptual connection via a new insight into contrastive learning methods. Noise-contrastive estimation can be used to optimize t-SNE, while UMAP relies on negative sampling, another contrastive method. We find the precise relationship between these two contrastive methods, and provide a mathematical characterization of the distortion introduced by negative sampling. Visually, this distortion results in UMAP generating more compact embeddings with tighter clusters compared to t-SNE. We exploit this new conceptual connection to propose and implement a generalization of negative sampling, allowing us to interpolate between (and even extrapolate beyond) t-SNE and UMAP and their respective embeddings. Moving along this spectrum of embeddings leads to a trade-off between discrete / local and continuous / global structures, mitigating the risk of over-interpreting ostensible features of any single embedding. We provide a PyTorch implementation.

****

On the optimal precision of GANs

Thibaut Issenhuth,Ugo Tanielian,Jeremie Mary,David Picard

Generative adversarial networks (GANs) are known to face model misspecification when learning disconnected distributions. Indeed, continuous mapping from a unimodal latent distribution to a disconnected one is impossible, so GANs necessarily generate samples outside of the support of the target distribution. In this paper, we make the connection between the performance of GANs and their latent space configuration. In particular, we raise the following question: what is the latent space partition that minimizes the measure of out-of-manifold samples? Building on a recent result of geometric measure theory, we prove there exist optimal GANs when the dimension of the latent space is larger than the number of modes. In particular, we show that these generators structure their latent space as a `simplicial cluster' - a Voronoi partition where centers are equally distant. We derive both an upper and a lower bound on the optimal precision of GANs learning disconnected manifolds. Interestingly, these two bounds have the same order of decrease: $\sqrt{\log m}$, m being the number of modes. Finally, we perform several experiments to exhibit the geometry of the latent space and experimentally show that GANs have a geometry with similar properties to the theoretical one.

****

D4AM: A General Denoising Framework for Downstream Acoustic Models

Chi-Chang Lee,Yu Tsao,Hsin-Min Wang,Chu-Song Chen

The performance of acoustic models degrades notably in noisy environments. Speech enhancement (SE) can be used as a front-end strategy to aid automatic speech recognition (ASR) systems. However, existing training objectives of SE methods are not fully effective at integrating speech-text and noise-clean paired data for training toward unseen ASR systems. In this study, we propose a general denoising framework, D4AM, for various downstream acoustic models. Our framework fine-tunes the SE model with the backward gradient according to a specific acoustic model and the corresponding classification objective. In addition, our method aims to consider the regression objective as an auxiliary loss to make the SE model generalize to other unseen acoustic models. To jointly train an SE unit with regression and classification objectives, D4AM uses an adjustment scheme to directly estimate suitable weighting coefficients rather than undergoing a grid search process with additional training costs. The adjustment scheme consists of two parts: gradient calibration and regression objective weighting. The experimental results show that D4AM can consistently and effectively provide improvements to various unseen acoustic models and outperforms other combination setups. Specifically, when evaluated on the Google ASR API with real noisy data completely unseen during SE training, D4AM achieves a relative WER reduction of 24.65% compared with the direct feeding of noisy input. To our knowledge, this is the first work that deploys an effective combination scheme of regression (denoising) and classification (ASR) objectives to derive a general pre-processor applicable to various unseen ASR systems. Our code is available at https://github.com/ChangLee0903/D4AM.

****

Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning

Qingru Zhang,Minshuo Chen,Alexander Bukharin,Pengcheng He,Yu Cheng,Weizhu Chen,Tuo Zhao

Fine-tuning large pre-trained language models on downstream tasks has become an important paradigm in NLP. However, common practice fine-tunes all of the parameters in a pre-trained model, which becomes prohibitive when a large number of downstream tasks are present. Therefore, many fine-tuning methods are proposed to learn incremental updates of pre-trained weights in a parameter efficient way, e.g., low-rank increments. These methods often evenly distribute the budget of incremental updates across all pre-trained weight matrices, and overlook the varying importance of different weight parameters. As a consequence, the fine-tuning performance is suboptimal. To bridge this gap, we propose AdaLoRA, which adaptively allocates the parameter budget among weight matrices according to their importance score. In particular, AdaLoRA parameterizes the incremental updates in the form of singular value decomposition. Such a novel approach allows us to effectively prune the singular values of unimportant updates, which is essentially to reduce their parameter budget but circumvent intensive exact SVD computations.

We conduct extensive experiments with several pre-trained models on natural language processing, question answering, and natural language generation to validate the effectiveness of AdaLoRA. Results demonstrate that AdaLoRA manifests notable improvement over baselines, especially in the low budget settings. Our code is publicly available at https://github.com/QingruZhang/AdaLoRA .

****

On Intriguing Layer-Wise Properties of Robust Overfitting in Adversarial Training

Duke Nguyen,Chaojian Yu,Vinoth Nandakumar,Young Lee,Tongliang Liu

Adversarial training has proven to be one of the most effective methods to defend against adversarial attacks. Nevertheless, robust overfitting is a common obstacle in adversarial training of deep networks. There is a common belief that the features learned by different network layers have different properties, however, existing works generally investigate robust overfitting by considering a DNN as a single unit and hence the impact of different network layers on robust overfitting remains unclear. In this work, we divide a DNN into a series of layers and investigate the effect of different network layers on robust overfitting. We find that different layers exhibit distinct properties towards robust overfitting, and in particular, robust overfitting is mostly related to the optimization of latter parts of the network. Based upon the observed effect, we propose a robust adversarial training (RAT) prototype: in a mini-batch, we optimize the front parts of the network as usual, and adopt additional measures to regularize the optimization of the latter parts. Based on the prototype, we designed two realizations of RAT, and extensive experiments demonstrate that RAT can eliminate robust overfitting and boost adversarial robustness over the standard adversarial training.

****

Does Federated Learning Really Need Backpropagation?

Haozhe Feng,Tianyu Pang,Chao Du,Wei Chen,Shuicheng YAN,Min Lin

Federated learning (FL) provides general principles for decentralized clients to train a server model collectively without sharing local data. FL is a promising framework with practical applications, but its standard training paradigm requires the clients to backpropagate through the model to compute gradients. Since these clients are typically edge devices and not fully trusted, executing backpropagation on them incurs computational and storage overhead as well as white-box vulnerability. In light of this, we develop backpropagation-free federated learning, dubbed BAFFLE, in which backpropagation is replaced by multiple forward processes to estimate gradients. BAFFLE is 1) memory-efficient and easily fits uploading bandwidth; 2) compatible with inference-only hardware optimization and model quantization or pruning; and 3) well-suited to trusted execution environments, because the clients in BAFFLE only execute forward propagation and return a set of scalars to the server. In experiments, we demonstrate that BAFFLE-trained models can achieve empirically comparable performance to conventional FL models.

****

Teaching Others is Teaching Yourself Regularization For Controllable Language Models

Han Liu,Bingning Wang,Ting Yao,Haijin Liang,Jianjin Xu,Xiaolin Hu

Large-scale pre-trained language models have achieved great success on natural language generation tasks. However, it is difficult to control the pre-trained language models to generate sentences with the expected attribute such as topic and sentiment. Recent efforts on controllable language generation employ an additional attribute classifier, which guides the generation of large-scale pre-trained language models, have been shown to be efficient in controllable language generation. These methods are named ''classifier-guided language models'' (CGLMs).

However, we find that the probabilities predicted by the attribute classifiers usually approaches 0 or 1, which make it hard to distinguish sentences with different matching degrees to the expected attribute. The problem is named \textit{the biased probability distribution} (BPD) problem.

To address the problem, we investigate different methods for adjusting probability distribution and propose a ''Teaching Others is Teaching Yourself'' (TOTY) re gularization method to smooth the probability distribution.

Experiments on sentiment control and topic control tasks show that CGLMs can get better performance with guiding classifiers trained with TOTY.

****

Specialization of Sub-paths for Adaptive Depth Networks

Woochul Kang

We present a novel approach to anytime networks that can control network depths instantly at runtime to provide various accuracy-efficiency trade-offs. While controlling the depth of a network is an effective way to obtain actual inference speed-up, previous adaptive depth networks require either additional intermediate classifiers or decision networks, that are challenging to train properly. Unlike previous approaches, our adaptive depth networks require virtually no architectural changes from baseline networks. Instead, we introduce a novel training method that enforces some sub-paths of the baseline networks to have a special property, with which the sub-paths do not change the semantic level of input features, but only refine them to reduce prediction errors. Those specialized sub-paths can be skipped at test time, if needed, to save computation at marginal loss of prediction accuracy. We first formally present the rationale behind the sub-paths specialization, and based on that, we propose a simple and practical training method to specialize sub-paths for adaptive depth networks. While minimal architectural changes and training efforts are required, we demonstrate that our approach significantly outperforms non-adaptive baselines in various tasks, including ImageNet classification, COCO object detection and instance segmentation. Further, we show that the smallest sub-networks of our adaptive depth networks achieve competitive model compression effect compared to recent state-of-the-art techniques.

****

Towards Effective and Interpretable Human-Agent Collaboration in MOBA Games: A Communication Perspective

Yiming Gao,Feiyu Liu,Liang Wang,Zhenjie Lian,Weixuan Wang,Siqin Li,Xianliang Wang,Xianhan Zeng,Rundong Wang,jiawei wang,QIANG FU,Yang Wei,Lanxiao Huang,Wei Liu

MOBA games, e.g., Dota2 and Honor of Kings, have been actively used as the testbed for the recent AI research on games, and various AI systems have been developed at the human level so far. However, these AI systems mainly focus on how to compete with humans, less on exploring how to collaborate with humans. To this end, this paper makes the first attempt to investigate human-agent collaboration in MOBA games. In this paper, we propose to enable humans and agents to collaborate through explicit communication by designing an efficient and interpretable Meta-Command Communication-based framework, dubbed MCC, for accomplishing effective human-agent collaboration in MOBA games. The MCC framework consists of two pivotal modules: 1) an interpretable communication protocol, i.e., the Meta-Command, to bridge the communication gap between humans and agents; 2) a meta-command value estimator, i.e., the Meta-Command Selector, to select a valuable meta-command for each agent to achieve effective human-agent collaboration. Experimental results in Honor of Kings demonstrate that MCC agents can collaborate reasonably well with human teammates and even generalize to collaborate with different levels and numbers of human teammates. Videos are available at https://sites.google.com/view/mcc-demo.

****

How Normalization and Weight Decay Can Affect SGD? Insights from a Simple Normalized Model

Ruosi Wan,Qiaosen Wang,Xiangyu Zhang,Yu-Wing Tai,Jian Sun,Chi-Keung Tang

Recent works(Li et al., 2020, Wan et al., 2021) characterize an important mechanism of normalized model trained with SGD and WD (Weight Decay), called Spherical Motion Dynamics (SMD), confirming its widespread effects in practice. However, no theoretical study is available on the influence of SMD on the training process of normalized models in literature. In this work, we seek to understand the effect of SMD by theoretically analyzing a simple normalized model, named as Noisy Rayleigh Quotient (NRQ). On NRQ, We theoretically prove SMD can dominate the whole training process via controlling the evolution of angular update (AU), an es sential feature of SMD. Specifically, we show: 1) within equilibrium state of SMD, the convergence rate and limiting risk of NRQ are mainly determined by the theoretical value of AU; and 2) beyond equilibrium state, the evolution of AU can interfere the optimization trajectory, causing odd phenomena such as ``escape'' behavior. We further show the insights drawn from NRQ is consistent with empirical observations in experiments on real datasets. We believe our theoretical results shed new light on the role of normalization techniques during the training of modern deep learning models.

****

Generalize Learned Heuristics to Solve Large-scale Vehicle Routing Problems in Real-time

Qingchun Hou,Jingwei Yang,Yiqiang Su,Xiaoqing Wang,Yuming Deng

Large-scale Vehicle Routing Problems (VRPs) are widely used in logistics, transportation, supply chain, and robotic systems. Recently, data-driven VRP heuristics are proposed to generate real-time VRP solutions with up to 100 nodes. Despite this progress, current heuristics for large-scale VRPs still face three major challenges: 1) Difficulty in generalizing the heuristics learned on small-scale VRPs to large-scale VRPs without retraining; 2) Challenge in generating real-time solutions for large-scale VRPs; 3) Difficulty in embedding global constraints into learned heuristics. We contribute in the three directions: We propose a Two-stage Divide Method (TAM) to generate sub-route sequence rather than node sequence for generalizing the heuristics learned on small-scale VRPs to solve large-scale VRPs in real-time. A two-step reinforcement learning method with new reward and padding techniques is proposed to train our TAM. A global mask function is proposed to keep the global constraints satisfied when dividing a large-scale VRP into several small-scale Traveling Salesman Problems (TSPs). As result, we can solve the small-scale TSPs in parallel quickly. The experiments on synthetic and real-world large-scale VRPs show our method could generalize the learned heuristics trained on datasets of VRP 100 to solve VRPs with over 5000 nodes in real-time while keeping the solution quality better than data-driven heuristics and competitive with traditional heuristics.

****

RetinexUTV: ROBUST RETINEX MODEL WITH UNFOLDING TOTAL VARIATION

Guiyu Guo,Daming shi,Zunjin Zhao,Muhammad Tahir Rasheed

Digital images are underexposed due to poor scene lighting or hardware limitations, reducing visibility and level of detail in the image, which will affect subsequent high-level tasks and image aesthetics. Therefore, it is of great practical significance to enhance low-light images. Among existing low-light image enhancement techniques, retinex-based methods are the focus today. However, most retinex methods either ignore or poorly handle noise during enhancement, which can produce unpleasant visual effects in low-light image enhancement and affect high-level tasks. In this paper, we propose a robust low-light image enhancement method RetinexUTV, which aims to enhance low-light images well while suppressing noise. In RetinexUTV, we propose an adaptive illumination estimation unfolded total variational network, which approximates the noise level of the real low-light image by learning the balance parameter of the total variation regularization term of the model, obtains the noise level map and the smooth noise-free sub-map of the image. The initial illumination map is then estimated by obtaining the illumination information of the smooth sub-map. The initial reflection map is obtained through the initial illumination map and original image. Under the guidance of the noise level map, the noise of the reflection map is suppressed, and finally it is multiplied by the adjusted illumination map to obtain the final enhancement result. We test our method on real low-light datasets LOL, VELOL, and experiments demonstrate that our method outperforms state-of-the-art methods.

****

Towards the Generalization of Contrastive Self-Supervised Learning

Weiran Huang,Mingyang Yi,Xuyang Zhao,Zihao Jiang

Recently, self-supervised learning has attracted great attention, since it only requires unlabeled data for model training. Contrastive learning is one popular method for self-supervised learning and has achieved promising empirical perform ance. However, the theoretical understanding of its generalization ability is still limited. To this end, we define a kind of (σ,δ)-measure to mathematically quantify the data augmentation, and then provide an upper bound of the downstream classification error rate based on the measure. It reveals that the generalization ability of contrastive self-supervised learning is related to three key factors: alignment of positive samples, divergence of class centers, and concentration of augmented data. The first two factors are properties of learned representations, while the third one is determined by pre-defined data augmentation. We further investigate two canonical contrastive losses, InfoNCE and cross-correlation, to show how they provably achieve the first two factors. Moreover, we conduct experiments to study the third factor, and observe a strong correlation between downstream performance and the concentration of augmented data.

****

Towards Controllable Policy through Goal-Masked Transformers

Xinyao Niu,Tong Sang,Yuchen Sun,Xiangjun Wang

Offline goal-conditioned supervised learning (GCSL) can learn to achieve various goals from purely offline datasets without reward information, enhancing control over the policy. However, we argue that learning a composite policy switchable among different goals seamlessly should be an essential task for obtaining a controllable policy. This feature should be learnable if the dataset contains enough data about such switches. Unfortunately, most existing datasets either partially or entirely lack such switching demonstrations. Current GCSL approaches that use hindsight information concentrate primarily on reachability at the state or return level. They might not work as expected when the goal is changed within an episode. To this end, we present Goal-Masked Transformers (GMT), an efficient GCSL algorithm based on transformers with goal masking. GMT makes use of trajectory-level hindsight information, which is automatically gathered and can be adjusted for various statistics of interest. Due to the autoregressive nature of GMT, we can change the goal and control the policy at any time. We empirically evaluate GMT on MuJoCo continuous control benchmarks and Atari discrete control games with image states to compare GMT against baselines. We illustrate that GMT can infer the missing switching processes from the given dataset and thus switch smoothly among different goals. As a result, GMT demonstrates its ability to control policy and succeeds on all the tasks with low variance, while existing GCSL works can hardly succeed in goal-switching.

****

Comparative Analysis between Vision Transformers and CNNs from the view of Neuroscience

Xiuyuan Hu,Hao Zhang,Yang Zhao

Neuroscience has provide many inspirations for the development of artificial intelligence, especially for neural networks for computer vision tasks. Recent research on animals' visual systems builds the connection between neural sparsity and animals' levels of evolution, based on which comparisons between two most influential vision architecture, Transformer and CNN, are carried out. In particular, the sparsity of attentions in Transformers is comprehensively studied, and previous knowledge on sparsity of neurons in CNNs is reviewed. In addition, a novel metric for neural sparsity is defined and ablation experiments are launched on various types of Transformer and CNN models. Finally, we draw the conclusion that more layers in models will result in higher sparsity, however, too many heads in Transformers may cause reduction of sparsity, which attributes to the significant overlap among effects of attention units.

****

Uncertainty-Aware Meta-Learning for Multimodal Task Distributions

Cesar Almecija,Apoorva Sharma,Navid Azizan

Meta-learning or learning to learn is a popular approach for learning new tasks with limited data (i.e., few-shot learning) by leveraging the commonalities among different tasks. However, meta-learned models can perform poorly when context data is limited, or when data is drawn from an out-of-distribution (OoD) task. Especially in safety-critical settings, this necessitates an uncertainty-aware ap proach to meta-learning. In addition, the often multimodal nature of task distributions can pose unique challenges to meta-learning methods. In this work, we present UnLiMTD (Uncertainty-aware meta-Learning for Multimodal Task Distributions), a novel method for meta-learning that (1) makes probabilistic predictions on in-distribution tasks efficiently, (2) is capable of detecting OoD context data at test time, and (3) performs on heterogeneous, multimodal task distributions. To achieve this goal, we take a probabilistic perspective and train a parametric, tuneable distribution over tasks on the meta-dataset. We construct this distribution by performing Bayesian inference on a linearized neural network, leveraging Gaussian process theory. We demonstrate that UnLiMTD’s predictions compare to, and outperform in most cases, the standard baselines, especially in the low-data regime. Furthermore, we show that UnLiMTD is effective in detecting data from OoD tasks. Finally, we confirm that both of these findings continue to hold in the multimodal task-distribution setting.

****

Neural Operator Variational Inference based on Regularized Stein Discrepancy for Deep Gaussian Processes

JIAN XU,Shian Du,Junmei Yang,Qianli Ma,Delu Zeng

A Deep Gaussian Process (DGP) model is a hierarchical composition of GP models that provides a deep Bayesian nonparametric approach to infer the posterior. Exact Bayesian inference is usually intractable for DGPs, motivating the use of various approximations. We theoretically demonstrate that the traditional alternative of mean-field Gaussian assumptions across the hierarchy leads to lack of expressiveness and efficacy of DGP models, whilst stochastic approximation often incurs a significant computational cost. To address this issue, we propose Neural Operator Variational Inference (NOVI) for Deep Gaussian Processes, where a sampler is obtained from a neural generator through minimizing Regularized Stein Discrepancy in L2 space between the approximate distribution and true posterior. Wherein a minimax problem is obtained and solved by Monte Carlo estimation and subsampling stochastic optimization. We experimentally demonstrate the effectiveness and efficiency of the proposed model, by applying it to a more flexible and wider class of posterior approximations on data ranging in size from hundreds to tens of thousands. By comparison, NOVI is superior to previous methods in both classification and regression.

****

On the complexity of nonsmooth automatic differentiation

Jerome Bolte,Ryan Boustany,Edouard Pauwels,Béatrice Pesquet-Popescu

Using the notion of conservative gradient, we provide a simple model to estimate the computational costs of the backward and forward modes of algorithmic differentiation for a wide class of nonsmooth programs. The complexity overhead of the backward mode turns out to be independent of the dimension when using programs with locally Lipschitz semi-algebraic or definable elementary functions. This extends considerably the Baur-Strassen's smooth cheap gradient principle. We illustrate our results by establishing fast backpropagation results of conservative gradients through feedforward neural networks with standard activation and loss functions. Nonsmooth backpropagation's cheapness contrasts with concurrent forward approaches, which have, to this day, dimensional-dependent worst case overhead estimates. We provide further results suggesting the superiority of backward propagation of conservative gradients. Indeed, we relate the complexity of computing a large number of directional derivatives to that of matrix multiplication, and we show that finding two subgradients in the Clarke subdifferential of a function is a NP-hard problem.

****

CO3: Cooperative Unsupervised 3D Representation Learning for Autonomous Driving

Runjian Chen,Yao Mu,Runsen Xu,Wenqi Shao,Chenhan Jiang,Hang Xu,Yu Qiao,Zhenguo Li,Ping Luo

Unsupervised contrastive learning for indoor-scene point clouds has achieved great successes. However, unsupervised representation learning on outdoor-scene point clouds remains challenging because previous methods need to reconstruct the whole scene and capture partial views for the contrastive objective. This is infe asible in outdoor scenes with moving objects, obstacles, and sensors. In this paper, we propose CO3, namely {Co}operative {Co}ntrastive Learning and {Co}ntextual Shape Prediction, to learn 3D representation for outdoor-scene point clouds in an unsupervised manner. CO3 has several merits compared to existing methods. (1) It utilizes LiDAR point clouds from vehicle-side and infrastructure-side to build views that differ enough but meanwhile maintain common semantic information for contrastive learning, which are more appropriate than views built by previous methods. (2) Alongside the contrastive objective, we propose contextual shape prediction to bring more task-relevant information for unsupervised 3D point cloud representation learning and we also provide a theoretical analysis for this pre-training goal. (3) As compared to previous methods, representation learned by CO3 is able to be transferred to different outdoor scene dataset collected by different type of LiDAR sensors. (4) CO3 improves current state-of-the-art methods on Once, KITTI and NuScenes datasets by up to 2.58 mAP in 3D object detection task and 3.54 mIoU in LiDAR semantic segmentation task. Codes and models will be released.

****

Bag of Tricks for Unsupervised Text-to-Speech

Yi Ren,Chen Zhang,Shuicheng YAN

Unsupervised text-to-speech (TTS) aims to train TTS models for a specific language without any paired speech-text training data in that language. Existing methods either use speech and corresponding pseudo text generated by an unsupervised automatic speech recognition (ASR) model as training data, or employ the back-translation technique. Though effective, they suffer from low robustness to low-quality data and heavy dependence on the lexicon of a language that is sometimes unavailable, leading to difficulty in convergence, especially in low-resource language scenarios. In this work, we introduce a bag of tricks to enable effective unsupervised TTS. Specifically, 1) we carefully design a voice conversion model to normalize the variable and noisy information in the low-quality speech data while preserving the pronunciation information; 2) we employ the non-autoregressive TTS model to overcome the robustness issue; and 3) we explore several tricks applied in back-translation, including curriculum learning, length augmentation and auxiliary supervised loss to stabilize the back-translation and improve its effectiveness. Through experiments, it has been demonstrated that our method achieves better intelligibility and audio quality than all previous methods, and that these tricks are very essential to the performance gain.

****

FedSpeed: Larger Local Interval, Less Communication Round, and Higher Generalization Accuracy

Yan Sun,Li Shen,Tiansheng Huang,Liang Ding,Dacheng Tao

Federated learning (FL) is an emerging distributed machine learning framework which jointly trains a global model via a large number of local devices with data privacy protections. Its performance suffers from the non-vanishing biases introduced by the local inconsistent optimal and the rugged client-drifts by the local over-fitting. In this paper, we propose a novel and practical method, FedSpeed, to alleviate the negative impacts posed by these problems. Concretely, FedSpeed applies the prox-correction term on the current local updates to efficiently reduce the biases introduced by the prox-term, a necessary regularizer to maintain the strong local consistency. Furthermore, FedSpeed merges the vanilla stochastic gradient with a perturbation computed from an extra gradient ascent step in the neighborhood, thereby alleviating the issue of local over-fitting. Our theoretical analysis indicates that the convergence rate is related to both the communication rounds T and local intervals K with a tighter upper bound $\mathcal{O}(\frac{1}{T})$ if $K=\mathcal{O}(T)$. Moreover, we conduct extensive experiments on the real-world dataset to demonstrate the efficiency of our proposed FedSpeed, which converges significantly faster and achieves the state-of-the-art (SOTA) performance on the general FL experimental settings than several baselines including FedAvg, FedProx, FedCM, FedAdam, SCAFFOLD, FedDyn, FedADMM, etc.

****

Holistically Explainable Vision Transformers

Moritz Böhle,Mario Fritz,Bernt Schiele

Transformers increasingly dominate the machine learning landscape across many tasks and domains, which increases the importance for understanding their outputs. While their attention modules provide partial insight into their inner workings, the attention scores have been shown to be insufficient for explaining the models as a whole. To address this, we propose B-cos transformers, which inherently provide holistic explanations for their decisions. Specifically, we formulate each model component—such as the multi-layer perceptrons, attention layers, and the tokenisation module—to be dynamic linear, which allows us to faithfully summarise the entire transformer via a single linear transform. We apply our proposed design to Vision Transformers (ViTs) and show that the resulting models, dubbed Bcos-ViTs, are highly interpretable and perform competitively to baseline ViTs on ImageNet. Code will be available at: github.com/anonymous/authors.

****

Neural Volumetric Mesh Generator

Yan Zheng,Lemeng Wu,Xingchao Liu,Zhen Chen,qiang liu,Qixing Huang

Deep generative models have shown success in generating 3D shapes with different representations. In this work, we propose Neural Volumetric Mesh Generator (NVMG), which can generate novel and high-quality volumetric meshes. Unlike the previous 3D generative model for point cloud, voxel, and implicit surface, the volumetric mesh representation is a ready-to-use representation in industry with details on both the surface and interior. Generating this such highly-structured data thus brings a significant challenge. We first propose a diffusion-based generative model to tackle this problem by generating voxelized shapes with close-to-reality outlines and structures. We can simply obtain a tetrahedral mesh as a template with the voxelized shape. Further, we use a voxel-conditional neural network to predict the smooth implicit surface conditioned on the voxels, and progressively project the tetrahedral mesh to the predicted surface under regularization. The regularization terms are carefully designed so that they can (1) get rid of the defects like flipping and high distortion; (2) force the regularity of the interior and surface structure during the deformation procedure for a high-quality final mesh. As shown in the experiments, our pipeline can generate high-quality artifact-free volumetric and surface meshes from random noise or a reference image without any post-processing. Compared with the state-of-the-art voxel-to-mesh deformation method, we show more robustness and better performance when taking generated voxels as input.

****

PathFusion: Path-consistent Lidar-Camera Deep Feature Fusion

Lemeng Wu,Dilin Wang,Meng Li,Yunyang Xiong,Raghuraman Krishnamoorthi,qiang liu,Vikas Chandra

Fusing camera with LiDAR is a promising technique to improve the accuracy of 3D detection due to its complementary physical properties.

While most existing methods focus on fusing camera features directly with raw LiDAR point clouds or shallow 3D features, it is observed that direct deep 3D feature fusion achieves inferior accuracy due to feature mis-alignment. The mis-alignment that originates from the feature aggregation across large receptive fields becomes increasingly severe for deep network stages. In this paper, we propose PathFusion to enable path-consistent LiDAR-camera deep feature fusion. PathFusion introduces a path consistency loss between shallow and deep features, which encourages the 2D backbone and its fusion path to transform 2D features in a way that is semantically aligned with the transform of the 3D backbone. We apply PathFusion to the prior-art fusion baseline, Focals Conv, and observe more than 1.2% mAP improvements on the nuScenes test split consistently with and without testing-time augmentations. Moreover, PathFusion also improves KITTI AP 3D (R11) by more than 0.6% on moderate level.

****

DADAO: Decoupled Accelerated Decentralized Asynchronous Optimization

Adel Nabli,Edouard Oyallon

DADAO is a novel decentralized asynchronous stochastic first order algorithm to minimize a sum of L-smooth and μ-strongly convex functions distributed ove r a time-varying connectivity network of size n. We model the local gradient updates and gossip communication procedures with separate independent Poisson Point Processes, decoupling the computation and communication steps in addition to making the whole approach completely asynchronous. Our method employs primal gradients and does not use a multi-consensus inner loop nor other ad-hoc mechanisms such as Error Feedback, Gradient Tracking, or a Proximal operator. By relating the inverse of the smallest positive eigenvalue χ^*_1 and the effective resistance χ_2^* of our graph to a necessary minimal communication rate between nodes of the network, we show that our algorithm requires $\mathcal{O}(n\sqrt{\frac{L}{\mu}}\log \epsilon)$ local gradients and only $\mathcal{O}(n\sqrt{\chi_1^*\chi_2^*}\sqrt{\frac{L}{\mu}}\log \epsilon)$ communications to reach a precision ϵ. If SGD with uniform noise σ^2 is used, we reach a precision ϵ with same speed, up to a bias term in $\mathcal{O}(\frac{\sigma^2}{\sqrt{\mu L}})$. This improves upon the bounds obtained with current state-of-the-art approaches, our simulations validating the strength of our relatively unconstrained method.

****

Advancing Radiograph Representation Learning with Masked Record Modeling

Hong-Yu Zhou,Chenyu Lian,Liansheng Wang,Yizhou Yu

Modern studies in radiograph representation learning (R^{2}L) rely on either self-supervision to encode invariant semantics or associated radiology reports to incorporate medical expertise, while the complementarity between them is barely noticed. To explore this, we formulate the self- and report-completion as two complementary objectives and present a unified framework based on masked record modeling (MRM). In practice, MRM reconstructs masked image patches and masked report tokens following a multi-task scheme to learn knowledge-enhanced semantic representations. With MRM pre-training, we obtain pre-trained models that can be well transferred to various radiography tasks. Specifically, we find that MRM offers superior performance in label-efficient fine-tuning. For instance, MRM achieves 88.5% mean AUC on CheXpert using 1% labeled data, outperforming previous R^{2}L methods with 100% labels. On NIH ChestX-ray, MRM outperforms the best performing counterpart by about 3% under small labeling ratios. Besides, MRM surpasses self- and report-supervised pre-training in identifying the pneumonia type and the pneumothorax area, sometimes by large margins.

****

Instance-wise Batch Label Restoration via Gradients in Federated Learning

Kailang Ma,Yu Sun,Jian Cui,Dawei Li,Zhenyu Guan,Jianwei Liu

Gradient inversion attacks have posed a serious threat to the privacy of federated learning. The attacks search for the optimal pair of input and label best matching the shared gradients and the search space of the attacks can be reduced by pre-restoring labels. Recently, label restoration technique allows for the extraction of labels from gradients analytically, but even the state-of-the-art remains limited to identify the presence of categories (i.e., the class-wise label restoration). This work considers the more real-world settings, where there are multiple instances of each class in a training batch. An analytic method is proposed to perform instance-wise batch label restoration from only the gradient of the final layer. On the basis of the approximate recovered class-wise embeddings and post-softmax probabilities, we establish linear equations of the gradients, probabilities and labels to derive the Number of Instances (NoI) per class by the Moore-Penrose pseudoinverse algorithm. Our experimental evaluations reach over 99% Label existence Accuracy (LeAcc) and exceed 96% Label number Accuracy (LnAcc) in most cases on three image datasets and four classification models. The two metrics are used to evaluate class-wise and instance-wise label restoration accuracy, respectively. And the recovery is made feasible even with a batch size of 4096 and partially negative activations (e.g., Leaky ReLU and Swish). Furthermore, we demonstrate that our method facilitates the existing gradient inversion attacks by exploiting the recovered labels, with an increase of 6-7 in PSNR on both MNIST and CIFAR100. Our code is
available at https://github.com/BUAA-CST/iLRG.

****

Re-parameterizing Your Optimizers rather than Architectures

Xiaohan Ding,Honghao Chen,Xiangyu Zhang,Kaiqi Huang,Jungong Han,Guiguang Ding

The well-designed structures in neural networks reflect the prior knowledge incorporated into the models. However, though different models have various priors, we are used to training them with model-agnostic optimizers such as SGD. In this paper, we propose to incorporate model-specific prior knowledge into optimizers by modifying the gradients according to a set of model-specific hyper-parameters. Such a methodology is referred to as Gradient Re-parameterization, and the optimizers are named RepOptimizers. For the extreme simplicity of model structure, we focus on a VGG-style plain model and showcase that such a simple model trained with a RepOptimizer, which is referred to as RepOpt-VGG, performs on par with or better than the recent well-designed models. From a practical perspective, RepOpt-VGG is a favorable base model because of its simple structure, high inference speed and training efficiency. Compared to Structural Re-parameterization, which adds priors into models via constructing extra training-time structures, RepOptimizers require no extra forward/backward computations and solve the problem of quantization. We hope to spark further research beyond the realms of model structure design. Code and models https://github.com/DingXiaoH/RepOptimizers.

****

Protein Representation Learning via Knowledge Enhanced Primary Structure Reasoning

Hong-Yu Zhou,Yunxiang Fu,Zhicheng Zhang,Bian Cheng,Yizhou Yu

Protein representation learning has primarily benefited from the remarkable development of language models (LMs). Accordingly, pre-trained protein models also suffer from a problem in LMs: a lack of factual knowledge. The recent solution models the relationships between protein and associated knowledge terms as the knowledge encoding objective. However, it fails to explore the relationships at a more granular level, i.e., the token level. To mitigate this, we propose Knowledge-exploited Auto-encoder for Protein (KeAP), which performs token-level knowledge graph exploration for protein representation learning. In practice, non-masked amino acids iteratively query the associated knowledge tokens to extract and integrate helpful information for restoring masked amino acids via attention. We show that KeAP can consistently outperform the previous counterpart on 9 representative downstream applications, sometimes surpassing it by large margins. These results suggest that KeAP provides an alternative yet effective way to perform knowledge enhanced protein representation learning.

****

The Provable Benefit of Unsupervised Data Sharing for Offline Reinforcement Learning

Hao Hu,Yiqin Yang,Qianchuan Zhao,Chongjie Zhang

Self-supervised methods have become crucial for advancing deep learning by leveraging data itself to reduce the need for expensive annotations. However, the question of how to conduct self-supervised offline reinforcement learning (RL) in a principled way remains unclear.

In this paper, we address this issue by investigating the theoretical benefits of utilizing reward-free data in linear Markov Decision Processes (MDPs) within a semi-supervised setting. Further, we propose a novel, Provable Data Sharing algorithm (PDS) to utilize such reward-free data for offline RL. PDS uses additional penalties on the reward function learned from labeled data to prevent overestimation, ensuring a conservative algorithm. Our results on various offline RL tasks demonstrate that PDS significantly improves the performance of offline RL algorithms with reward-free data. Overall, our work provides a promising approach to leveraging the benefits of unlabeled data in offline RL while maintaining theoretical guarantees. We believe our findings will contribute to developing more robust self-supervised RL methods.

****

Federated Learning for Inference at Anytime and Anywhere

Zicheng Liu,Da Li,Javier Fernandez-Marques,Stefanos Laskaridis,Yan Gao,■ukasz Dudziak,Stan Z. Li,Shell Xu Hu,Timothy Hospedales

Federated learning has been predominantly concerned with collaborative training of deep networks from scratch, and especially the many challenges that arise, such as communication cost, robustness to heterogeneous data, and support for diverse device capabilities. However, there is no unified framework that addresses all these problems together. This paper studies the challenges and opportunities of exploiting pre-trained Transformer models in FL. In particular, we propose to efficiently adapt such pre-trained models by injecting a novel attention-based adapter module at each transformer block that both modulates the forward pass and makes an early prediction. Training only the lightweight adapter by FL leads to fast and communication-efficient learning even in the presence of heterogeneous data and devices. Extensive experiments on standard FL benchmarks, including CIFAR- 100, FEMNIST and SpeechCommandsv2 demonstrate that this simple framework provides fast and accurate FL while supporting heterogenous device capabilities, efficient personalization, and scalable-cost anytime inference.
****

Modeling Sequential Sentence Relation to Improve Cross-lingual Dense Retrieval
Shunyu Zhang,Yaobo Liang,MING GONG,Daxin Jiang,Nan Duan
Recently multi-lingual pre-trained language models (PLM) such as mBERT and XLM-R have achieved impressive strides in cross-lingual dense retrieval. Despite its successes, they are general-purpose PLM while the multilingual PLM tailored for cross-lingual retrieval is still unexplored. Motivated by an observation that the sentences in parallel documents are approximately in the same order, which is universal across languages, we propose to model this sequential sentence relation to facilitate cross-lingual representation learning. Specifically, we propose a multilingual PLM called masked sentence model (MSM), which consists of a sentence encoder to generate the sentence representations, and a document encoder applied to a sequence of sentence vectors from a document. The document encoder is shared for all languages to model the universal sequential sentence relation across languages. To train the model, we propose a masked sentence prediction task, which masks and predicts the sentence vector via a hierarchical contrastive loss with sampled negatives. Comprehensive experiments on four cross-lingual retrieval tasks show MSM significantly outperforms existing advanced pre-training models, demonstrating the effectiveness and stronger cross-lingual retrieval capabilities of our approach.
****

A Robustly and Effectively Optimized Pretraining Approach for Masked Autoencoder
Ruijia Xu,Yixiao Ge,Kun Yi,XUYUAN XU,Yexin Wang,Ying-Cong Chen,Hao Chen,Ying Shan
Recently, Masked Image Modeling (MIM) has increasingly reshaped the status quo of self-supervised visual pre-training. This paper does not describe a novel MIM method, but to unravel several fundamental ingredients to robustly and effectively pre-train a Masked AutoEncoder (MAE) with improved downstream performance as a byproduct. We highlight the great significance for the whole autoencoder to encourage high-variance interactions across different tokens, while simultaneously for the reconstructed target to smooth the inter-patch variances. First, at the decoding phase, we apply the standard dropout upon the attention probabilities as noise to randomly mask out the edge connection across different tokens. Otherwise, their shortcut interactions might hinder the emergence of meaningful contextual representation. Second, we point out that the per-patch normalization will fail unless the patch pixels rely on some population statistics to reduce inter-patch variance and then smooth the reconstruction. Third, we show that autoencoders with different capacities encounter the issue to varying degrees and the learnable masked tokens can be employed to manipulate the variance dependent on its inserted position and ratio in the model. The proposed techniques here are simple and effective to benefit the pre-training of a masked autoencoder stably and obtain superior performance across different downstream tasks.
****

Diffusion Posterior Sampling for General Noisy Inverse Problems
Hyungjin Chung,Jeongsol Kim,Michael Thompson Mccann,Marc Louis Klasky,Jong Chul Ye

Diffusion models have been recently studied as powerful generative inverse problem solvers, owing to their high quality reconstructions and the ease of combining existing iterative solvers. However, most works focus on solving simple linear inverse problems in noiseless settings, which significantly under-represents the complexity of real-world problems. In this work, we extend diffusion solvers to efficiently handle general noisy (non)linear inverse problems via the Laplace approximation of the posterior sampling. Interestingly, the resulting posterior sampling scheme is a blended version of diffusion sampling with the manifold constrained gradient without a strict measurement consistency projection step, yielding a more desirable generative path in noisy settings compared to the previous studies. Our method demonstrates that diffusion models can incorporate various measurement noise statistics such as Gaussian and Poisson, and also efficiently handle noisy nonlinear inverse problems such as Fourier phase retrieval and non-uniform deblurring.

****

Low-Rank Graph Neural Networks Inspired by the Weak-balance Theory in Social Networks

Langzhang Liang,Xiangjing Hu,Zenglin Xu,Zixing Song,Irwin King

Graph Neural Networks (GNNs) have achieved state-of-the-art performance on node classification tasks by exploiting both the graph structures and node features. Generally, most existing GNNs depend on the implicit homophily assumption that nodes belonging to the same class are more likely to be connected. However, GNNs may fail to model heterophilious graphs where nodes with different labels tend to be linked, as shown in recent studies. To address this issue, we propose a generic GNN applicable to both homophilious and heterophilious graphs, namely Low-Rank Graph Neural Network (LRGNN). In detail, we aim at computing a coefficient matrix such that the sign of each coefficient reveals whether the corresponding two nodes belong to the same class, which is similar to the sign inference problem. In Signed Social Networks (SSNs), the sign inference problem can be modeled as a low-rank matrix factorization (LRMF) problem due to the global low-rank structure described by the weak balance theory. In this paper, we show that signed graphs are naturally generalized weakly-balanced when considering node classification tasks. Motivated by this observation, we propose to leverage LRMF to recover a coefficient matrix from a partially observed signed adjacency matrix. To effectively capture the node similarity, we further incorporate the low-rank representation (LRR) method. Our theoretical result shows that under our update rule of node representations, LRR obtained by solving a subspace clustering problem can recover the subspace structure of node representations. To solve the corresponding optimization problem, we utilize an iterative optimization algorithm with a convergence guarantee and develop a neural-style initialization manner that enables fast convergence. Finally, extensive experimental evaluation on both real-world and synthetic graphs has validated the superior performance of LRGNN over various state-of-the-art GNNs. In particular, LRGNN can offer clear performance gains in a scenario when the node features are not informative enough.

****

Do We Need Neural Collapse? Learning Diverse Features for Fine-grained and Long-tail Classification

Jiawei Ma,Chong You,Sashank J. Reddi,Sadeep Jayasumana,Himanshu Jain,Felix Yu,Shih-Fu Chang,Sanjiv Kumar

Feature extractors learned from supervised training of deep neural networks have demonstrated superior performance over handcrafted ones. Recently, it is shown that such learned features have a neural collapse property, where within-class features collapse to the class mean and different class means are maximally separated. This paper examines the neural collapse property in the context of fine-grained classification tasks, where a feature extractor pretrained from a classification task with coarse labels is used for generating features for a downstream classification task with fine-grained labels. We argue that the within-class feature collapse is an undesirable property for fine-grained classification. Hence, we introduce a geometric arrangement of features called the maximal-separating-cone, where within-class features lie in a cone of nontrivial radius instead of collapsing to the class mean, and cones of different classes are maximally separated. We present a technique based on classifier weight and training loss design to produce such an arrangement. Experimentally we demonstrate an improved fine-grained classification performance with a feature extractor pretrained by our method. Moreover, our technique also provides benefits for the classification on data with long-tail distribution over classes. Our work may motivate future efforts on the design of better geometric arrangements of deep features.

****

HRBP: Hardware-friendly Regrouping towards Block-wise Pruning for Sparse Training

Haoyu Ma,Chengming Zhang,lizhi xiang,Xiaolong Ma,Geng Yuan,Wenkai Zhang,Shiwei Liu,Tianlong Chen,Dingwen Tao,Yanzhi Wang,Zhangyang Wang,Xiaohui Xie

Recently, pruning at initialization and training a sparse network from scratch (sparse training) become increasingly popular. However, most sparse training literature addresses only the unstructured sparsity, which in practice brings little benefit to the training acceleration on GPU due to the irregularity of non-zero weights. In this paper, we work on sparse training with fine-grained structured sparsity, by extracting a few dense blocks from unstructured sparse weights. For Convolutional Neural networks (CNN), however, the extracted dense blocks will be broken in backpropagation due to the shape transformation of convolution filters implemented by GEMM. Thus, previous block-wise pruning methods can only be used to accelerate the forward pass of sparse CNN training. To this end, we propose the Hardware-friendly Regrouping towards Block-based Pruning (HRBP), where the grouping is conducted on the kernel-wise mask. With HRBP, extracted dense blocks are preserved in backpropagation. We further propose HRBP++ to reduce zero kernels by extracting common sparse kernel patterns on all kernels within one block. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet demonstrate that HRBP (HRBP++) can almost match the accuracy of unstructured sparse training methods while achieving a huge acceleration on hardware.

****

DepthFL : Depthwise Federated Learning for Heterogeneous Clients

Minjae Kim,Sangyoon Yu,Suhyun Kim,Soo-Mook Moon

Federated learning is for training a global model without collecting private local data from clients. As they repeatedly need to upload locally-updated weights or gradients instead, clients require both computation and communication resources enough to participate in learning, but in reality their resources are heterogeneous. To enable resource-constrained clients to train smaller local models, width scaling techniques have been used, which reduces the channels of a global model. Unfortunately, width scaling suffers from heterogeneity of local models when averaging them, leading to a lower accuracy than when simply excluding resource-constrained clients from training. This paper proposes a new approach based on depth scaling called DepthFL. DepthFL defines local models of different depths by pruning the deepest layers off the global model, and allocates them to clients depending on their available resources. Since many clients do not have enough resources to train deep local models, this would make deep layers partially-trained with insufficient data, unlike shallow layers that are fully trained. DepthFL alleviates this problem by mutual self-distillation of knowledge among the classifiers of various depths within a local model. Our experiments show that depth-scaled local models build a global model better than width-scaled ones, and that self-distillation is highly effective in training data-insufficient deep layers.

****

Masked Image Modeling with Denoising Contrast

Kun Yi,Yixiao Ge,Xiaotong Li,Shusheng Yang,Dian Li,Jianping Wu,Ying Shan,Xiaohu Qie

Since the development of self-supervised visual representation learning from contrastive learning to masked image modeling (MIM), there is no significant difference in essence, that is, how to design proper pretext tasks for vision dictionary look-up. MIM recently dominates this line of research with state-of-the-art performance on vision Transformers (ViTs), where the core is to enhance the patch

-level visual context capturing of the network via denoising auto-encoding mechanism. Rather than tailoring image tokenizers with extra training stages as in previous works, we unleash the great potential of contrastive learning on de- noising auto-encoding and introduce a pure MIM method, ConMIM, to produce simple intra-image inter-patch contrastive constraints as the sole learning objectives for masked patch prediction. We further strengthen the denoising mechanism with asymmetric designs, including image perturbations and model progress rates, to improve the network pre-training. ConMIM-pretrained models with various scales achieve competitive results on downstream image classification, semantic segmentation, object detection, and instance segmentation tasks, e.g., on ImageNet-1K classification, we achieve 83.9% top-1 accuracy with ViT-Small and 85.3% with ViT-Base without extra data for pre-training. Code will be available at https://github.com/TencentARC/ConMIM.

****

Monkeypox with Cross Infection Hypothesis via Epidemiological Mode
Tahir Khan,zhiqiang xu
A new re-emerging infectious disease of monkeypox 2022 is structurally related to smallpox that is induced by the monkeypox viruses and has caused 59,606 active cases with 18 deaths up to September 15, 2022. To end this ongoing epidemic, there is a need for population-wide control policies like reducing social interaction by keeping social distance, treatment of infected individuals, and restriction on animals, etc. We forecast the progression of the epidemic and come up with an efficient control mechanism by formulating a mathematical model. The biological feasibility and dynamical behavior of the proposed model are then investigated together with sensitivity analysis to obtain the effect of various epidemic parameters mitigating the spread of the disease. Subsequently, by taking non-pharmaceutical and pharmaceutical intervention strategies as control measures, an optimal control theory is applied to mitigate the fatality of the disease to minimize the infectious population and reduce the cost of controls, we construct an objective functional and solve it by using Pontryagin’s maximum principle. Finally, extensive numerical simulations are performed to show the impact of the application of intervention mechanisms in controlling the transmission of the monkeypox epidemic.

****

LPMARL: Linear Programming based Implicit Task Assignment for Hierarchical Multi-agent Reinforcement Learning
Kyuree Ahn,Jinkyoo Park
Training a multi-agent reinforcement learning (MARL) model with sparse reward is notoriously difficult because the terminal reward is induced by numerous interactions among agents. In this study, we propose linear programming-based hierarchical MARL (LPMARL) to learn effective coperative strategy among agents. LPMARL is composed of two hierarchical decision-making schemes: (1) solving an agent-task assignment problem and (2) solving a local cooperative game among agents that are assigned to the same task. For the first step, LPMARL formulates the agent-task assignment problem as linear programming (LP) using the state-dependent cost parameters generated by a graph neural network (GNN). Solving the LP can be considered as assigning tasks to agents, which decomposes the original problem into a set of task-dependent sub-problems. After solving the formulated LP, LPMARL employs a general MARL strategy to derive a lower-level policy to solve each sub-task in a cooperative manner. We train the LP-parameter generating GNN layer and the low-level MARL policy network, which are the essential components for making hierarchical decisions, in an end-to-end manner using the implicit function theorem. We empirically demonstrate that LPMARL learns an optimal agent-task allocation and the subsequent local cooperative control policy among agents in sub-groups for solving various mixed cooperative-competitive environments.

****

Transmission Dynamics of Hepatitis B: Analysis and Control
Tahir Khan,zhiqiang xu,Xin Cao
The infection of hepatitis B attacks the liver and can produce acute and chronic

diseases, while it is a major health problem and life-threatening around the globe. The control of this infection is a difficult task due to several reasons such as variation of human behavior, proper medication, vaccination, and existence of a large number of carries, etc., but understanding the dynamics of the infection helps to design appropriate control strategies. Thus, a proper complex dynamical system is needed to find the stability conditions and propose intervention strategies for forecasting the control of hepatitis B virus transmission. We formulate a model that will be helpful to investigate the temporal dynamics and suggest control strategies for hepatitis B infection. The well-posedness of the proposed model will be shown, and used to find the threshold parameter to analyze the model equilibria and its stability. We also perform the sensitive analysis of the threshold quantity to quantify the most sensitive epidemic parameters. Based on the temporal dynamics and sensitivity, we investigate effective methods to minimize the infection of hepatitis B, and develop the algorithms to support the theoretical results with the help of numerical simulations.
****

Mass-Editing Memory in a Transformer

Kevin Meng,Arnab Sen Sharma,Alex J Andonian,Yonatan Belinkov,David Bau

Recent work has shown exciting promise in updating large language models with new memories, so as to replace obsolete information or add specialized knowledge. However, this line of work is predominantly limited to updating single associations. We develop MEMIT, a method for directly updating a language model with many memories, demonstrating experimentally that it can scale up to thousands of associations for GPT-J (6B) and GPT-NeoX (20B), exceeding prior work by an order of magnitude. Our code and data will be open-sourced upon publication.
****

Enhancement and Numerical Assessment of Novel SARS-CoV-2 Virus Transmission Model

Tahir Khan,zhiqiang xu

Recent pandemic of the coronavirus started in December 2019, which has affected almost all groups of humankind. In this regard, accurate epidemic models are not only crucial for demonstrating the mitigation of the current pandemic but also helpful for forecasting their future dynamics. In this work, we propose a model for SARS-CoV-2 virus transmission to forecast the temporal dynamics of the novel coronavirus disease by considering the characteristics of the disease and the recent literature. Due to the nondeterministic and stochastic nature of the novel-coronavirus disease, we present the model with the aid of stochastic differential equations by considering two infectious phases: pre-symptomatic and symptomatic, because both are significant in the spread of SARS-CoV-2 virus transmission. We ensure that the model is well-posed and identify the necessary conditions for disease eradication by proving the existence, uniqueness, and extinction analysis. The efficacy of the model and the importance of the current study are demonstrated using the actual data. Finally, the model will be simulated using Euler-Maruyama and Milstein’s numerical schemes to support the theoretical findings and show the significance of the results obtained.
****

GoBigger: A Scalable Platform for Cooperative-Competitive Multi-Agent Interactive Simulation

Ming Zhang,Shenghan Zhang,Zhenjie Yang,Lekai Chen,Jinliang Zheng,Chao Yang,Chuming Li,Hang Zhou,Yazhe Niu,Yu Liu

The emergence of various multi-agent environments has motivated powerful algorithms to explore agents' cooperation or competition. Even though this has greatly promoted the development of multi-agent reinforcement learning (MARL), it is still not enough to support further exploration on the behavior of swarm intelligence between multiple teams, and cooperation between multiple agents due to their limited scalability. To alleviate this, we introduce GoBigger, a scalable platform for cooperative-competition multi-agent interactive simulation. GoBigger is an enhanced environment for the Agar-like game, enabling the simulation of multiple scales of agent intra-team cooperation and inter-team competition. Compared with existing multi-agent simulation environments, our platform supports multi-t eam games with more than two teams simultaneously, which dramatically expands the diversity of agent cooperation and competition, and can more effectively simulate the swarm intelligent agent behavior. Besides, in GoBigger, the cooperation between the agents in a team can lead to much higher performance. We offer a diverse set of challenging scenarios, built-in bots, and visualization tools for best practices in benchmarking. We evaluate several state-of-the-art algorithms on GoBigger and demonstrate the potential of the environment. We believe this platform can inspire various emerging research directions in MARL, swarm intelligence, and large-scale agent interactive learning. Both GoBigger and its related benchmark are open-sourced. More information could be found at https://github.com/opendilab/GoBigger.

****

Masked Unsupervised Self-training for Label-free Image Classification

Junnan Li,Silvio Savarese,Steven Hoi

State-of-the-art computer vision models are mostly trained with supervised learning using human-labeled images, which limits their scalability due to the expensive annotation cost. While self-supervised representation learning has achieved impressive progress, it still requires a second stage of finetuning on labeled data. On the other hand, models pre-trained with large-scale text supervision (e.g., CLIP) have enabled zero-shot transfer to downstream image classification tasks. However, the zero-shot performance of CLIP-like models are often insufficient for real-world adoption. In this paper, we aim to leverage the abundant unlabeled data from a target domain to improve the performance of a pre-trained zero-shot classifier, by unsupervised finetuning of the pre-trained model. We propose Masked Unsupervised Self-Training (MUST), a new approach which leverages two different and complimentary sources of training signals: pseudo-labels and raw images. MUST jointly optimizes three objectives to learn both class-level global feature and pixel-level local feature and enforces a regularization between the two. We demonstrate the efficacy of MUST on 8 downstream tasks across a variety of domains, where it improves upon CLIP by a large margin. MUST also outperforms supervised few-shot adaptation methods. It achieves a top-1 accuracy of 77.7% on ImageNet using ViT-B, +9.4% higher than CLIP, and +6.2% higher than 16-shot CLIP adaptation. Our code is available at https://github.com/salesforce/MUST.

****

Recursion of Thought: Divide and Conquer Reasoning with Language Models

Soochan Lee,Gunhee Kim

With the recent advances in language models, attempts are being made to apply them to solving multi-step reasoning problems. A major breakthrough in this line of research is to let language models generate intermediate steps, often called Chain of Thought (CoT), before producing a final answer. However, language models have an upper bound on the context size, i.e., the number of input tokens, such as 2048 for the recent GPT-3 and PaLM. Although several thousand tokens are enough to handle various tasks, solving more complex reasoning tasks can require orders of magnitude more tokens. Therefore, the context limit imposes a fundamental limit on the model's reasoning capability. Inspired by human's incredible reasoning ability based on abstraction and recursion, we propose Recursion of Thought (RoT) as a model-agnostic framework with the novel paradigm of teaching a language model to divide and conquer complex problems by recursively creating multiple contexts. Since RoT casts the context-related operations as tokens, a language model can trigger the recursion operations by simply producing the corresponding tokens. On multiple arithmetic and algorithmic reasoning tasks, we demonstrate that RoT dramatically improves the recent large-scale language model GPT-3 to solve extremely complex problems. Moreover, RoT can make tiny, randomly initialized Transformers or LSTMs to solve problems that even humans find daunting.

****

GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis

Zhenhui Ye,Ziyue Jiang,Yi Ren,Jinglin Liu,Jinzheng He,Zhou Zhao

Generating photo-realistic video portraits with arbitrary speech audio is a crucial problem in film-making and virtual reality. Recently, several works explore the usage of neural radiance field (NeRF) in this task to improve 3D realness an d image fidelity. However, the generalizability of previous NeRF-based methods is limited by the small scale of training data. In this work, we propose GeneFace, a generalized and high-fidelity NeRF-based talking face generation method, which can generate natural results corresponding to various out-of-domain audio. Specifically, we learn a variational motion generator on a large lip-reading corpus, and introduce a domain adaptative post-net to calibrate the result. Moreover, we learn a NeRF-based renderer conditioned on the predicted motion. A head-aware torso-NeRF is proposed to eliminate the head-torso separation problem. Extensive experiments show that our method achieves more generalized and high-fidelity talking face generation compared to previous methods. Video samples and source code are available at https://geneface.github.io .

****

Learning the Positions in CountSketch

Yi Li,Honghao Lin,Simin Liu,Ali Vakilian,David Woodruff

We consider sketching algorithms which first compress data by multiplication with a random sketch matrix, and then apply the sketch to quickly solve an optimization problem, e.g., low-rank approximation and regression. In the learning-based sketching paradigm proposed by Indyk et al., the sketch matrix is found by choosing a random sparse matrix, e.g., CountSketch, and then the values of its non-zero entries are updated by running gradient descent on a training data set. Despite the growing body of work on this paradigm, a noticeable omission is that the locations of the non-zero entries of previous algorithms were fixed, and only their values were learned.

In this work, we propose the first learning-based algorithms that also optimize the locations of the non-zero entries. Our first proposed algorithm is based on a greedy algorithm. However, one drawback of the greedy algorithm is its slower training time. We fix this issue and propose approaches for learning a sketching matrix for both low-rank approximation and Hessian approximation for second-order optimization. The latter is helpful for a range of constrained optimization problems, such as LASSO and matrix estimation with a nuclear norm constraint. Both approaches achieve good accuracy with a fast running time. Moreover, our experiments suggest that our algorithm can still reduce the error significantly even if we only have a very limited number of training matrices.

****

Towards the gradient adjustment by loss status for Neural Network Optimization

Jiexin Wang,Wenwen Qiang,Bing Su

Gradient descent-based algorithms are crucial in neural network optimization, and most of them only depend on local properties such as the first and second-order momentum of gradients to determine the local optimization directions. As a result, such algorithms often converge slowly in the case of a small gradient and easily fall into the local optimum. Since the goal of optimization is to minimize the loss function, the status of the loss indicates the overall progress of the optimization but has not been fully explored. In this paper, we propose a loss-aware gradient adjusting strategy (LGA) based on the loss status. LGA automatically adjusts the update magnitude of parameters to accelerate convergence and escape local optimums by introducing a loss-incentive correction term monitoring the loss and adapting gradient experience. The proposed strategy can be applied to various gradient descent-based optimization algorithms. We provide theoretical analysis on the convergence rate and empirical evaluations on different datasets to demonstrate the effectiveness of our method.

****

On the Necessity of Disentangled Representations for Downstream Tasks

Ruiqian Nai,Zixin Wen,Ji Li,Yuanzhi Li,Yang Gao

A disentangled representation encodes generative factors of data in a separable and compact pattern. Thus it is widely believed that such a representation format benefits downstream tasks. In this paper, we challenge the necessity of disentangled representation in downstream applications. Specifically, we show that dimension-wise disentangled representations are not necessary for downstream tasks using neural networks that take learned representations as input. We provide extensive empirical evidence against the necessity of disentanglement, covering mu ltiple datasets, representation learning methods, and downstream network architectures. Moreover, our study reveals that the informativeness of representations best accounts for downstream performance. The positive correlation between informativeness and disentanglement explains the claimed usefulness of disentangled representations in previous works.

****

Linear Video Transformer with Feature Fixation

Kaiyue Lu,ZeXiang Liu,Jianyuan Wang,Weixuan Sun,Zhen Qin,Dong Li,Xuyang Shen,Hui Deng,Xiaodong Han,Yuchao Dai,Yiran Zhong

Vision Transformers have achieved impressive performance in video classification, while suffering from the quadratic complexity caused by the Softmax attention mechanism. Some studies alleviate the computational costs by reducing the number of tokens attended in attention calculation, but the complexity is still quadratic. Another promising way is to replace Softmax attention with linear attention, which owns linear complexity but presents a clear performance drop. We find that such a drop in linear attention results from the lack of attention concentration to critical features. Therefore, we propose a feature fixation module to reweight feature importance of the query and key prior to computing linear attention. Specifically, we regard the query, key, and value as latent representations of the input token, and learn the feature fixation ratio by aggregating Query-Key-Value information. This is beneficial for measuring the feature importance comprehensively. Furthermore, we improve the feature fixation by neighborhood association, which leverages additional guidance from spatial and temporal neighboring tokens. Our proposed method significantly improves the linear attention baseline, and achieves state-of-the-art performance among linear video Transformers on three popular video classification benchmarks. Our performance is even comparable to some quadratic Transformers with fewer parameters and higher efficiency.

****

Neural Frailty Machine: Beyond proportional hazard assumption in neural survival regressions

Jiawei Qiao,Ruofan Wu,Mingzhe Wu,Wen Yu,Ming Zheng,Tengfei LIU,Tianyi Zhang,Weiqiang Wang

We present neural frailty machine (NFM), a powerful and flexible neural modeling framework for survival regressions. The NFM framework utilizes the classical idea of multiplicative frailty in survival analysis to capture unobserved heterogeneity among individuals, at the same time being able to leverage the strong approximation power of neural architectures for handling nonlinear covariate dependence. Two concrete models are derived under the framework that extends neural proportional hazard models and nonparametric hazard regression models. Both models allow efficient training under the likelihood objective. Theoretically, for both proposed models, we establish statistical guarantees of neural function approximation with respect to nonparametric components via characterizing their rate of convergence. Empirically, we provide synthetic experiments that verify our theoretical statements. We also conduct experimental evaluations over 6 benchmark datasets of different scales, showing that the proposed NFM models outperform state-of-the-art survival models in terms of predictive performance.

****

A Closer Look at Dual Batch Normalization and Two-domain Hypothesis In Adversarial Training With Hybrid Samples

Chaoning Zhang,Kang Zhang,Chenshuang Zhang,Axi Niu,Chang D. Yoo,In So Kweon

There is a growing concern about applying batch normalization (BN) in adversarial training (AT), especially when the model is trained on both \textit{adversarial} samples and \textit{clean} samples (termed Hybrid-AT). With the assumption that \textit{adversarial} and \textit{clean} samples are from two different domains, a common practice in prior works is to adopt dual BN, where BN$_{adv}$ and BN$_{clean}$ are used for adversarial and clean branches, respectively. A popular belief for motivating dual BN is that estimating normalization statistics of this mixture distribution is challenging and thus disentangling it for normalization achieves stronger robustness. In contrast to this belief, we reveal that what makes dual BN effective mainly lies in its two sets of affine parameters. Moreov

er, we demonstrate that the domain gap between adversarial and clean samples is actually not very large, which is counter-intuitive considering the significant influence of adversarial perturbation on the model. Overall, our work sheds new light on understanding the mechanism of dual BN in Hybrid-AT as well as its underlying two-domain hypothesis.
****

Generative Recorrupted-to-Recorrupted: An Unsupervised Image Denoising Network for Arbitrary Noise Distribution
wan Bowen,Daming Shi

With the great breakthrough of supervised learning in the field of denoising, more and more works focus on end-to-end learning to train denoisers. The premise of this method is effective is that there is certain data support, but in practice, it is particularly difficult to obtain labels in the training data. To this end, some unsupervised denoisers have emerged in recent years, however, the premise of these methods being effective is that the noise model needs to be known in advance, which will limit the practical use of unsupervised denoising. In addition, inaccurate noise prior from noise estimation algorithms causes low denoising accuracy. Therefore, we design a more practical denoiser that requires neither clean images as training labels nor noise model assumptions. Our method also needs the support of the noise model, the difference is that the model is generated by a residual image and a random mask during the network training process, and then the input and target of the network are generated from a single noisy images and the noise model, at the same time, train an unsupervised module and a pseudo supervised module. Extensive experiments demonstrate the effectiveness of our framework and even surpass the accuracy of supervised denoising.
****

Provably Learning Diverse Features in Multi-View Data with Midpoint Mixup
Muthu Chidambaram,Xiang Wang,Chenwei Wu,Rong Ge

Mixup is a data augmentation technique that relies on training using random convex combinations of data points and their labels. In recent years, Mixup has become a standard primitive used in the training of state-of-the-art image classification models due to its demonstrated benefits over empirical risk minimization with regards to generalization and robustness. In this work, we try to explain some of this success from a feature learning perspective. We focus our attention on classification problems in which each class may have multiple associated features (or views) that can be used to predict the class correctly. Our main theoretical results demonstrate that, for a non-trivial class of data distributions with two features per class, training a 2-layer convolutional network using empirical risk minimization can lead to learning only one feature for almost all classes while training with a specific instantiation of Mixup succeeds in learning both features for every class. We also show empirically that these theoretical insights extend to the practical settings of image benchmarks modified to have additional synthetic features.
****

Understanding Catastrophic Overfitting in Fast Adversarial Training From a Non-robust Feature Perspective
Chaoning Zhang,Chenshuang Zhang,Kang Zhang,Axi Niu,In So Kweon,Chang D. Yoo

To make adversarial training (AT) computationally efficient, FGSM AT has attracted significant attention. The fast speed, however, is achieved at the cost of catastrophic overfitting (CO), whose reason remains unclear. Prior works mainly study the phenomenon of a significant PGD accuracy (Acc) drop to understand CO while paying less attention to its FGSM Acc. We highlight an intriguing CO phenomenon that FGSM Acc is higher than accuracy on clean samples and attempt to apply non-robust feature (NRF) to understand it. Our investigation of CO by extending the existing NRF into fine-grained categorization suggests: there exists a certain type of NRF whose usefulness is increased after FGSM attack, and CO in FGSM AT can be seen as a dynamic process of learning such NRF. Therefore, the key to preventing CO lies in reducing its usefulness under FGSM AT, which sheds new light on understanding the success of a SOTA technique for mitigating CO.

****

BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers

Zhiliang Peng,Li Dong,Hangbo Bao,Qixiang Ye,Furu Wei

Masked image modeling (MIM) has demonstrated impressive results in self-supervised representation learning by recovering corrupted image patches. However, most existing studies operate on low-level image pixels, which hinders the exploitation of high-level semantics for representation models. In this work, we propose to use a semantic-rich visual tokenizer as the reconstruction target for masked prediction, providing a systematic way to promote MIM from pixel-level to semantic-level. Specifically, we propose vector-quantized knowledge distillation to train the tokenizer, which discretizes a continuous semantic space to compact codes. We then pretrain vision Transformers by predicting the original visual tokens for the masked image patches. Furthermore, we introduce a patch aggregation strategy which associates discrete image patches to enhance global semantic representation. Experiments on image classification and semantic segmentation show that BEiT v2 outperforms all compared MIM methods. On ImageNet-1K (224 size), the base-size BEiT v2 achieves 85.5% top-1 accuracy for fine-tuning and 80.1% top-1 accuracy for linear probing. The large-size BEiT v2 obtains 87.3% top-1 accuracy for ImageNet-1K (224 size) fine-tuning, and 56.7% mIoU on ADE20K for semantic segmentation. The code can be found in the supplementary materials.

****

HIVE: HIerarchical Volume Encoding for Neural Implicit Surface Reconstruction

Xiaodong Gu,Weihao Yuan,Heng Li,Zilong Dong,Ping Tan

Neural implicit surface reconstruction has become a new trend in reconstructing a detailed 3D shape from images. In previous methods, however, the 3D scene is only encoded by the MLPs which do not have an explicit 3D structure. To better represent 3D shapes, we introduce a volume encoding to explicitly encode the spatial information. We further design hierarchical volumes to encode the scene structures in multiple scales. The high-resolution volumes capture the high-frequency geometry details since spatially varying features could be learned from different 3D points, while the low-resolution volumes enforce the spatial consistency to keep the shape smooth since adjacent locations possess the same low-resolution feature. In addition, we adopt a sparse structure to reduce the memory consumption at high-resolution volumes, and two regularization terms to enhance results smoothness. This hierarchical volume encoding could be appended to any implicit surface reconstruction method as a plug-and-play module, and can generate a smooth and clean reconstruction with more details. Superior performance is demonstrated in DTU, EPFL, and BlendedMVS datasets with significant improvement on the standard metrics. The code of our method will be made public.

****

Communication-Efficient Federated Learning with Accelerated Client Gradient

Geeho Kim,Jinkyu Kim,Bohyung Han

Federated learning often suffers from slow and unstable convergence due to heterogeneous characteristics of participating client datasets.

Such a tendency is aggravated when the client participation ratio is low since the information collected from the clients is prone to have large variations.

To tackle this challenge, we propose a novel federated learning framework, which improves the consistency across clients and facilitates the convergence of the server model.

This is achieved by making the server broadcast a global model with a gradient acceleration.

By adopting the strategy, the proposed algorithm conveys the projective global update information to participants effectively with no extra communication cost and relieves the clients from storing the previous models.

We also regularize local updates by aligning each of the clients with the overshot global model to reduce bias and improve the stability of our algorithm.

We perform comprehensive empirical studies on real data under various settings and demonstrate remarkable performance gains of the proposed method in terms of accuracy and communication efficiency compared to the state-of-the-art methods, especially with low client participation rates.

We will release our code to facilitate and disseminate our work.

****

DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

Hao Zhang,Feng Li,Shilong Liu,Lei Zhang,Hang Su,Jun Zhu,Lionel Ni,Heung-Yeung Shum

We present DINO (DETR with Improved deNoising anchOr boxes), a strong end-to-end object detector. DINO improves over previous DETR-like models in performance and efficiency by using a contrastive way for denoising training, a look forward twice scheme for box prediction, and a mixed query selection method for anchor initialization. DINO achieves 49.4AP in 12 epochs and 51.3AP in 24 epochs on COCO with a ResNet-50 backbone and multi-scale features, yielding a significant improvement of +6.0AP and +2.7AP, respectively, compared to DN-DETR, the previous best DETR-like model. DINO scales well in both model size and data size. Without bells and whistles, after pre-training on the Objects365 dataset with a SwinL backbone, DINO obtains the best results on both COCO val2017 (63.2AP) and test-dev (63.3AP) with model size under 1 billion parameters. Compared to other models on the leaderboard, DINO significantly reduces its model size and pre-training data size while achieving better results. The code will be available.

****

Simultaneously Learning Stochastic and Adversarial Markov Decision Process with Linear Function Approximation

Fang Kong,XiangCheng Zhang,Baoxiang Wang,Shuai Li

Reinforcement learning (RL) has been commonly used in practice. To deal with the numerous states and actions in real applications, the function approximation method has been widely employed to improve the learning efficiency, among which the linear function approximation has attracted great interest both theoretically and empirically. Previous works on the linear Markov Decision Process (MDP) mainly study two settings, the stochastic setting where the reward is generated in a stochastic way and the adversarial setting where the reward can be chosen arbitrarily by an adversary. All these works treat these two environments separately. However, the learning agents often have no idea of how rewards are generated and a wrong reward type can severely disrupt the performance of those specially designed algorithms. So a natural question is whether an algorithm can be derived that can efficiently learn in both environments but without knowing the reward type. In this paper, we first consider such best-of-both-worlds problem for linear MDP with the known transition. We propose an algorithm and prove it can simultaneously achieve $O(\text{poly} \log K)$ regret in the stochastic setting and $O(\sqrt{K})$ regret in the adversarial setting where $K$ is the horizon. To the best of our knowledge, it is the first such result for linear MDP.

****

Statistical Efficiency of Score Matching: The View from Isoperimetry

Frederic Koehler,Alexander Heckett,Andrej Risteski

Deep generative models parametrized up to a normalizing constant (e.g. energy-based models) are difficult to train by maximizing the likelihood of the data because the likelihood and/or gradients thereof cannot be explicitly or efficiently written down. Score matching is a training method, whereby instead of fitting the likelihood $\log p(x)$ for the training data, we instead fit the score function $\nabla_x \log p(x)$ --- obviating the need to evaluate the partition function. Though this estimator is known to be consistent, its unclear whether (and when) its statistical efficiency is comparable to that of maximum likelihood --- which is known to be (asymptotically) optimal. We initiate this line of inquiry in this paper, and show a tight connection between statistical efficiency of score matching and the isoperimetric properties of the distribution being estimated --- i.e. the Poincar\'e, log-Sobolev and isoperimetric constant --- quantities which govern the mixing time of Markov processes like Langevin dynamics. Roughly, we show that the score matching estimator is statistically comparable to the maximum likelihood when the distribution has a small isoperimetric constant. Conversely, if the distribution has a large isoperimetric constant --- even for simple families of distributions like exponential families with rich enough sufficient statistics --- score matching will be substantially less efficient than maxim um likelihood. We suitably formalize these results both in the finite sample regime, and in the asymptotic regime. Finally, we identify a direct parallel in the discrete setting, where we connect the statistical properties of pseudolikelihood estimation with approximate tensorization of entropy and the Glauber dynamics.

****
Quadratic models for understanding neural network dynamics
Libin Zhu,Chaoyue Liu,Adityanarayanan Radhakrishnan,Misha Belkin

In this work, we show that recently proposed quadratic models capture optimization and generalization properties of wide neural networks that cannot be captured by linear models. In particular, we prove that quadratic models for shallow ReLU networks exhibit the "catapult phase" from 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.

****
Revisiting Graph Adversarial Attack and Defense From a Data Distribution Perspective
Kuan Li,Yang Liu,Xiang Ao,Qing He

Recent studies have shown that structural perturbations are significantly effective in degrading the accuracy of Graph Neural Networks (GNNs) in the semi-supervised node classification (SSNC) task. However, why the gradient-based methods are so destructive is rarely explored. In this work, we discover an interesting phenomenon: the adversarial edges are not uniformly distributed on the graph. Nearly all perturbations are generated around the training nodes in poisoning attack. Combined with this phenomenon, we provide an explanation for the effectiveness of the gradient-based attack method from a data distribution perspective and revisit both poisoning attack and evasion attack in SSNC. From this new perspective, we empirically and theoretically discuss some other attack tendencies. Based on the analysis, we provide nine practical tips on both attack and defense and meanwhile leverage them to improve existing attack and defense methods. Moreover, we design a fast attack method and a self-training defense method, which outperform the state-of-the-art methods and can effectively scale to large graphs like ogbn-arxiv. We conduct extensive experiments on four benchmark datasets to verify our claims.

****
Gated Domain Units for Multi-source Domain Generalization
Simon Föll,Alina Dubatovka,Eugen Ernst,Martin Maritsch,Patrik Okanovic,Gudrun Thaeter,Joachim M. Buhmann,Felix Wortmann,Krikamol Muandet

Distribution shift (DS) is a common problem that deteriorates the performance of learning machines. To tackle this problem, we postulate that real-world distributions are composed of elementary distributions that remain invariant across different environments. We call this an invariant elementary distribution (I.E.D.) assumption. The I.E.D. assumption implies an invariant structure in the solution space that enables knowledge transfer to unseen domains. To exploit this property in domain generalization (DG), we developed a modular neural network layer that consists of Gated Domain Units (GDUs). Each GDU learns an embedding of an individual elementary distribution that allows us to encode the domain similarities during the training. During inference, the GDUs compute similarities between an observation and each of the corresponding elementary distributions which are then used to form a weighted ensemble of learning machines. Because our layer is trained with backpropagation, it can naturally be integrated into existing deep learning frameworks. Our evaluation on image, text, graph, and time-series data shows a significant improvement in the performance on out-of-training target domains without domain information and any access to data from the target domains. This finding supports the practicality of the I.E.D. assumption and demonstrates that our GDUs can learn to represent these elementary distributions.

****

Learning large-scale Kernel Networks

Amirhesam Abedsoltan,Parthe Pandit,Misha Belkin

This paper concerns large-scale training of *Kernel Networks*, a generalization of kernel machines that allows the model to have arbitrary centers. We propose a scalable training algorithm -- EigenPro 3.0 -- based on alternating projections with preconditioned SGD for the alternating steps. In contrast to classical kernel machines, but similar to neural networks, our algorithm enables decoupling the learned model from the training set. This empowers kernel models to take advantage of modern methodologies in deep learning, such as data augmentation. We demonstrate the promise of EigenPro 3.0 on several experiments over large datasets. We also show data augmentation can improve performance of kernel models.

****

Provable Sim-to-real Transfer in Continuous Domain with Partial Observations

Jiachen Hu,Han Zhong,Chi Jin,Liwei Wang

Sim-to-real transfer, which trains RL agents in the simulated environments and then deploys them in the real world, has been widely used to overcome the limitations of gathering samples in the real world. Despite the empirical success of the sim-to-real transfer, its theoretical foundation is much less understood. In this paper, we study the sim-to-real transfer in continuous domain with partial observations, where the simulated environments and real-world environments are modeled by linear quadratic Gaussian (LQG) systems. We show that a popular robust adversarial training algorithm is capable of learning a policy from the simulated environment that is competitive to the optimal policy in the real-world environment. To achieve our results, we design a new algorithm for infinite-horizon average-cost LQGs and establish a regret bound that depends on the intrinsic complexity of the model class. Our algorithm crucially relies on a novel history clipping scheme, which might be of independent interest.

****

Local Coefficient Optimization in Federated Learning

Congliang Chen,Chi Zhang,Bingzhe Wu,Yatao Bian,Zhi-Quan Luo,Peilin Zhao

Federated learning emerges as a promising approach to build a large-scale cooperative learning system among multiple clients without sharing their raw data. However, given a specific global objective, finding the optimal sampling weights for each client remains largely unexplored. This is particularly challenging when clients' data distributions are non-i.i.d. and clients partially participate.

In this paper, we model the above task as a bi-level optimization problem which takes the correlations among different clients into account. We present a double-loop primal-dual-based algorithm to solve the bi-level optimization problem. We further provide rigorous convergence analysis for our algorithm under mild assumptions. Finally, we perform extensive empirical studies under both toy examples and learning models from real datasets to verify the effectiveness of the proposed method.

****

Outcome-directed Reinforcement Learning by Uncertainty \& Temporal Distance-Aware Curriculum Goal Generation

Daesol Cho,Seungjae Lee,H. Jin Kim

Current reinforcement learning (RL) often suffers when solving a challenging exploration problem where the desired outcomes or high rewards are rarely observed. Even though curriculum RL, a framework that solves complex tasks by proposing a sequence of surrogate tasks, shows reasonable results, most of the previous works still have difficulty in proposing curriculum due to the absence of a mechanism for obtaining calibrated guidance to the desired outcome state without any prior domain knowledge. To alleviate it, we propose an uncertainty \& temporal distance-aware curriculum goal generation method for the outcome-directed RL via solving a bipartite matching problem. It could not only provide precisely calibrated guidance of the curriculum to the desired outcome states but also bring much better sample efficiency and geometry-agnostic curriculum goal proposal capability compared to previous curriculum RL methods. We demonstrate that our algorithm significantly outperforms these prior methods in a variety of challenging navigation tasks and robotic manipulation tasks in a quantitative and qualitative way.

****

Synergistic Neuromorphic Federated Learning with ANN-SNN Conversion For Privacy Protection
Yingni Chen,Shikuang Deng,Yuhang Li,Xin Dong,Shi Gu

Federated Learning (FL) has been widely explored for the growing public data privacy issues, where only model parameters are communicated instead of private data. However, recent studies debunk the privacy protection of FL, showing that private data can be leaked from the communicated gradients or parameters updates. In this paper, we propose a framework called Synergistic Neuromorphic Federated Learning (SNFL) that enhances privacy during FL. Before uploading the updates of the client model, SNFL first converts clients' Artificial Neural Networks (ANNs) to Spiking Neural Networks (SNNs) via calibration algorithms. In a way that not only loses almost no accuracy but also encrypts the client model's parameters, SNFL manages to obtain a more performant model with high privacy. After aggregation of various SNNs parameters, server distributes the parameters back to clients to continue training under ANN architecture, providing smooth convergence. The proposed framework is demonstrated to be private, introducing a lightweight overhead as well as yielding prominent performance boosts. Extensive experiments with different kinds of datasets have demonstrated the efficacy and practicability of our method. In most of our experimental IID and not extreme Non-IID scenarios, the SNFL technique has significantly enhanced the model performance. For instance, SNFL improve the accuracy of FedAvg on Tiny-ImageNet by 13.79%. In the IID situation of tiny-ImageNet, for instance, the SNFL method is 13.79% more accurate than FedAvg. Also, the original image cannot be reconstructed after 280 iterations of attacks with the SNFL method, whereas it can be reconstructed after just 70 iterations with FedAvg.

****

Federated Semi-supervised Learning with Dual Regulator
Sikai Bai,Shuaicheng Li,Weiming Zhuang,Kunlin Yang,Jun Hou,Shuai Yi,Shuai Zhang,Junyu Gao,Song Guo

Federated learning emerges as a powerful method to learn from decentralized heterogeneous data while protecting data privacy. Federated semi-supervised learning (FSSL) is even more practical and challenging, where only a fraction of data can be labeled due to high annotation cost. Existing FSSL methods, however, assume independent and identically distributed (IID) labeled data across clients and consistent class distribution between labeled and unlabeled data within a client. In this work, we propose a novel FSSL framework with dual regulator, FedDure, to optimize and customize model training according to specific data distributions of clients. FedDure lifts the previous assumption with a coarse-grained regulator (C-reg) and a fine-grained regulator (F-reg): C-reg regularizes the updating of local model by tracking the learning effect on labeled data distribution; F-reg learns an adaptive weighting scheme tailored for unlabeled instances in each client. We further formulate the client model training as bi-level optimization that adaptively optimize the model in the client with two regulators. Theoretically, we show the convergence guarantee of dual regulator. Empirically, we demonstrate that FedDure is superior to the existing methods across wide range of settings, notably by more than 12% on CIFAR-10 and CINIC-10 datasets.

****

Path Regularization: A Convexity and Sparsity Inducing Regularization for Parallel ReLU Networks
Tolga Ergen,Mert Pilanci

Understanding the fundamental principles behind the success of deep neural networks is one of the most important open questions in the current literature. To this end, we study the training problem of deep neural networks and introduce an analytic approach to unveil hidden convexity in the optimization landscape. We consider a deep parallel ReLU network architecture, which also includes standard d eep networks and ResNets as its special cases. We then show that pathwise regularized training problems can be represented as an exact convex optimization problem. We further prove that the equivalent convex problem is regularized via a group sparsity inducing norm. Thus, a path regularized parallel ReLU network can be viewed as a parsimonious convex model in high dimensions. More importantly, we show that the computational complexity required to globally optimize the equivalent convex problem is fully polynomial-time in feature dimension and number of samples. Therefore, we prove polynomial-time trainability of path regularized ReLU networks with global optimality guarantees. We also provide experiments corroborating our theory.

****

Globally Optimal Training of Neural Networks with Threshold Activation Functions

Tolga Ergen,Halil Ibrahim Gulluk,Jonathan Lacotte,Mert Pilanci

Threshold activation functions are highly preferable in neural networks due to their efficiency in hardware implementations. Moreover, their mode of operation is more interpretable and resembles that of biological neurons. However, traditional gradient based algorithms such as Gradient Descent cannot be used to train the parameters of neural networks with threshold activations since the activation function has zero gradient except at a single non-differentiable point. To this end, we study weight decay regularized training problems of deep neural networks with threshold activations. We first show that regularized deep threshold network training problems can be equivalently formulated as a standard convex optimization problem, which parallels the LASSO method, provided that the last hidden layer width exceeds a certain threshold. We also derive a simplified convex optimization formulation when the dataset can be shattered at a certain layer of the network. We corroborate our theoretical results with various numerical experiments.

****

Molecule Generation For Target Protein Binding with Structural Motifs

ZAIXI ZHANG,Yaosen Min,Shuxin Zheng,Qi Liu

Designing ligand molecules that bind to specific protein binding sites is a fundamental problem in structure-based drug design. Although deep generative models and geometric deep learning have made great progress in drug design, existing works either sample in the 2D graph space or fail to generate valid molecules with realistic substructures. To tackle these problems, we propose a Fragment-based LigAnd Generation framework (FLAG), to generate 3D molecules with valid and realistic substructures fragment-by-fragment. In FLAG, a motif vocabulary is constructed by extracting common molecular fragments (i.e., motif) in the dataset. At each generation step, a 3D graph neural network is first employed to encode the intermediate context information. Then, our model selects the focal motif, predicts the next motif type, and attaches the new motif. The bond lengths/angles can be quickly and accurately determined by cheminformatics tools. Finally, the molecular geometry is further adjusted according to the predicted rotation angle and the structure refinement. Our model not only achieves competitive performances on conventional metrics such as binding affinity, QED, and SA, but also outperforms baselines by a large margin in generating molecules with realistic substructures.

****

Bag of Tricks for FGSM Adversarial Training

Zichao Li,Li Liu,Zeyu Wang,Yuyin Zhou,Cihang Xie

Adversarial training (AT) with samples generated by Fast Gradient Sign Method (FGSM), also known as FGSM-AT, is a computationally simple method to train robust networks. However, during its training procedure, an unstable mode of ``catastrophic overfitting'' has been identified in [Wong2020FastIB], where the robust accuracy abruptly drops to zero within a single training step. Existing methods use gradient regularizers or random initialization tricks to attenuate this issue, whereas they either take high computational cost or lead to lower robust accuracy. In this work, we provide the first study, which thoroughly examines a collection of tricks from three perspectives: Data Initialization, Network Structure, and Optimization, to overcome the catastrophic overfitting in FGSM-AT.

Surprisingly, we find that simple tricks, i.e., a) masking partial pixels (even without randomness), b) setting a large convolution stride and smooth activation functions, or c) regularizing the weights of the first convolutional layer, can effectively tackle the overfitting issue.

Extensive results on a range of network architectures validate the effectiveness of each proposed trick, and the combinations of tricks are also investigated. For example, trained with PreActResNet-18 on CIFAR-10, our method attains 49.8% accuracy against PGD-50 attacker and 46.4% accuracy against AutoAttack, demonstrating that pure FGSM-AT is capable of enabling robust learners.

****

Towards Robustness Certification Against Universal Perturbations

Yi Zeng,Zhouxing Shi,Ming Jin,Feiyang Kang,Lingjuan Lyu,Cho-Jui Hsieh,Ruoxi Jia

In this paper, we investigate the problem of certifying neural network robustness against universal perturbations (UPs), which have been widely used in universal adversarial attacks and backdoor attacks. Existing robustness certification methods aim to provide robustness guarantees for each sample with respect to the worst-case perturbations given a neural network. However, those sample-wise bounds will be loose when considering the UP threat model as they overlook the important constraint that the perturbation should be shared across all samples. We propose a method based on a combination of linear relaxation-based perturbation analysis and Mixed Integer Linear Programming to establish the first robust certification method for UP. In addition, we develop a theoretical framework for computing error bounds on the entire population using the certification results from a randomly sampled batch. Aside from an extensive evaluation of the proposed certification, we further show how the certification facilitates efficient comparison of robustness among different models or efficacy among different universal adversarial attack defenses and enables accurate detection of backdoor target classes.

****

Deep Generative Modeling on Limited Data with Regularization by Nontransferable Pre-trained Models

Yong Zhong,Hong Tao Liu,Xiaodong Liu,Fan Bao,Weiran Shen,Chongxuan Li

Deep generative models (DGMs) are data-eager because learning a complex model on limited data suffers from a large variance and easily overfits. Inspired by the classical perspective of the bias-variance tradeoff, we propose regularized deep generative model (Reg-DGM), which leverages a nontransferable pre-trained model to reduce the variance of generative modeling with limited data. Formally, Reg-DGM optimizes a weighted sum of a certain divergence and the expectation of an energy function, where the divergence is between the data and the model distributions, and the energy function is defined by the pre-trained model w.r.t. the model distribution. We analyze a simple yet representative Gaussian-fitting case to demonstrate how the weighting hyperparameter trades off the bias and the variance. Theoretically, we characterize the existence and the uniqueness of the global minimum of Reg-DGM in a non-parametric setting and prove its convergence with neural networks trained by gradient-based methods. Empirically, with various pre-trained feature extractors and a data-dependent energy function, Reg-DGM consistently improves the generation performance of strong DGMs with limited data and achieves competitive results to the state-of-the-art methods. Our implementation is available at https://github.com/ML-GSAI/Reg-ADA-APA.

****

Defense against Backdoor Attacks via Identifying and Purifying Bad Neurons

Mingyuan Fan,Yang Liu,Cen Chen,Ximeng Liu,Wenzhong Guo

Recent studies reveal the vulnerability of neural networks to backdoor attacks. By embedding backdoors into the hidden neurons with poisoned training data, the backdoor attacker can override normal predictions of the victim model to the attacker-chosen ones whenever the backdoor pattern is present in a testing input. In this paper, to mitigate public concerns about the attack, we propose a novel backdoor defense via identifying and purifying the backdoored neurons of the victim neural network. Specifically, we first define a new metric, called benign sal ience. By combining the first-order gradient to retain the connections between neurons, benign salience can identify the backdoored neurons with high accuracy. Then, a new Adaptive Regularization (AR) mechanism is proposed to assist in purifying these identified bad neurons via fine-tuning. Due to the ability to adapt to different magnitudes of parameters, AR can provide faster and more stable convergence than the common regularization mechanisms in neuron purifying. Finally, we test the defense effect of our method on ten different backdoor attacks with three benchmark datasets. Experimental results show that our method can decrease the attack success rate by more than 95% on average, which is the best among six state-of-the-art defense methods.

****

Basic Binary Convolution Unit for Binarized Image Restoration Network

Bin Xia,Yulun Zhang,Yitong Wang,Yapeng Tian,Wenming Yang,Radu Timofte,Luc Van Gool

Lighter and faster image restoration (IR) models are crucial for the deployment on resource-limited devices. Binary neural network (BNN), one of the most promising model compression methods, can dramatically reduce the computations and parameters of full-precision convolutional neural networks (CNN). However, there are different properties between BNN and full-precision CNN, and we can hardly use the experience of designing CNN to develop BNN. In this study, we reconsider components in binary convolution, such as residual connection, BatchNorm, activation function, and structure, for IR tasks. We conduct systematic analyses to explain each component's role in binary convolution and discuss the pitfalls. Specifically, we find that residual connection can reduce the information loss caused by binarization; BatchNorm can solve the value range gap between residual connection and binary convolution; The position of the activation function dramatically affects the performance of BNN. Based on our findings and analyses, we design a simple yet efficient basic binary convolution unit (BBCU). Furthermore, we divide IR networks into four parts and specially design variants of BBCU for each part to explore the benefit of binarizing these parts. We conduct experiments on different IR tasks, and our BBCU significantly outperforms other BNNs and lightweight models, which shows that BBCU can serve as a basic unit for binarized IR networks. All codes and models will be released.

****

DSP: Dynamic Semantic Prototype for Generative Zero-Shot Learning

Shiming Chen,Hou Wen Jin,Ziming Hong,Yibing Song,Tongliang Liu,Xinge You,Kun Zhang

Generative models (e.g., generative adversarial network (GAN)) have advanced zero-shot learning (ZSL). Studies on the generative ZSL methods typically produce visual features of unseen classes to mitigate the issue of lacking unseen samples based on the predefined class semantic prototypes. As these empirically designed prototypes are not able to faithfully represent the actual semantic prototypes of visual features (i.e., visual prototypes), existing methods limit their ability to synthesize visual features that accurately represent real features and prototypes. We formulate this phenomenon as a visual-semantic domain shift problem. It prevents the generative models from further improving the ZSL performance. In this paper, we propose a dynamic semantic prototype learning (DSP) method to align the empirical and actual semantic prototypes for synthesizing accurate visual features. The alignment is conducted by jointly refining semantic prototypes and visual features so that the generator synthesizes visual features which are close to the real ones. We utilize a visual$\rightarrow$semantic mapping network (V2SM) to map both the synthesized and real features into the class semantic space. The V2SM benefits the generator to synthesize visual representations with rich semantics. The real/synthesized visual features supervise our visual-oriented semantic prototype evolving network (VOPE) where the predefined class semantic prototypes are iteratively evolved to become dynamic semantic prototypes. Such prototypes are then fed back to the generative network as conditional supervision. Finally, we enhance visual features by fusing the evolved semantic prototypes into their corresponding visual features. Our extensive experiments on three benchmark datasets show that our DSP improves existing generative ZSL methods, \t extit{e.g.}, the average improvements of the harmonic mean over four baselines (e.g., CLSWGAN, f-VAEGAN, TF-VAEGAN and FREE) by 8.5\%, 8.0\% and 9.7\% on CUB, SUN and AWA2, respectively.

****

Analyzing the Latent Space of GAN through Local Dimension Estimation

Jaewoong Choi,Geonho Hwang,Hyunsoo Cho,Myungjoo Kang

The impressive success of style-based GANs (StyleGANs) in high-fidelity image synthesis has motivated research to understand the semantic properties of their latent spaces. Recently, a close relationship was observed between the semantically disentangled local perturbations and the local PCA components in the learned latent space $\mathcal{W}$. However, understanding the number of disentangled perturbations remains challenging. Building upon this observation, we propose a local dimension estimation algorithm for an arbitrary intermediate layer in a pre-trained GAN model. The estimated intrinsic dimension corresponds to the number of disentangled local perturbations. In this perspective, we analyze the intermediate layers of the mapping network in StyleGANs. Our analysis clarifies the success of $\mathcal{W}$-space in StyleGAN and suggests a method for finding an alternative. Moreover, the intrinsic dimension estimation opens the possibility of unsupervised evaluation of global-basis-compatibility and disentanglement for a latent space. Our proposed metric, called Distortion, measures an inconsistency of intrinsic tangent space on the learned latent space. The metric is purely geometric and does not require any additional attribute information. Nevertheless, the metric shows a high correlation with the global-basis-compatibility and supervised disentanglement score. Our findings pave the way towards an unsupervised selection of globally disentangled latent space among the intermediate latent spaces in a GAN.

****

A Causal Approach to Detecting Multivariate Time-series Anomalies and Root Causes

Wenzhuo Yang,Kun Zhang,Steven Hoi

Detecting anomalies and the corresponding root causes in multivariate time series plays an important role in monitoring the behaviors of various real-world systems, e.g., IT system operations or manufacturing industry. Previous anomaly detection approaches model the joint distribution without considering the underlying mechanism of multivariate time series, making them computationally hungry and hard to identify root causes. In this paper, we formulate the anomaly detection problem from a causal perspective and view anomalies as instances that do not follow the regular causal mechanism to generate the multivariate data. We then propose a causality-based framework for detecting anomalies and root causes. It first learns the causal structure from data and then infers whether an instance is an anomaly relative to the local causal mechanism whose conditional distribution can be directly estimated from data. In light of the modularity property of causal systems (the causal processes to generate different variables are irrelevant modules), the original problem is divided into a series of separate, simpler, and low-dimensional anomaly detection problems so that where an anomaly happens (root causes) can be directly identified. We evaluate our approach with both simulated and public datasets as well as a case study on real-world AIOps applications, showing its efficacy, robustness, and practical feasibility.

****

Quark: A Gradient-Free Quantum Learning Framework for Classification Tasks

Zhihao Zhang,Zhuoming Chen,Heyang Huang,Zhihao Jia

As more practical and scalable quantum computers emerge, much attention has been focused on realizing quantum supremacy in machine learning. Existing quantum ML methods either (1) embed a classical model into a target Hamiltonian to enable quantum optimization or (2) represent a quantum model using variational quantum circuits and apply classical gradient-based optimization. The former method leverages the power of quantum optimization but only supports simple ML models, while the latter provides flexibility in model design but relies on gradient calculation, resulting in barren plateau (i.e., gradient vanishing) and frequent classical-quantum interactions. To address the limitations of existing quantum ML meth

ods, we introduce Quark, a gradient-free quantum learning framework that optimizes quantum ML models using quantum optimization. Quark does not rely on gradient computation and therefore avoids barren plateau and frequent classical-quantum interactions. In addition, Quark can support more general ML models than prior quantum ML methods and achieves a dataset-size-independent optimization complexity. Theoretically, we prove that Quark can outperform classical gradient-based methods by reducing model query complexity for highly non-convex problems; empirically, evaluations on the Edge Detection and Tiny-MNIST tasks show that Quark can support complex ML models and significantly reduce the number of measurements needed for discovering near-optimal weights for these tasks.
****

Cross-modal Graph Contrastive Learning with Cellular Images
Shuangjia Zheng,Jiahua Rao,Jixian Zhang,Cohen Ethan,Chengtao Li,Yuedong Yang
Constructing discriminative representations of molecules lies at the core of a number of domains such as drug discovery, material science, and chemistry. State-of-the-art methods employ graph neural networks (GNNs) and self-supervised learning (SSL) to learn the structural representations from unlabeled data, which can then be fine-tuned for downstream tasks. Albeit powerful, these methods that are pre-trained solely on molecular structures cannot generalize well to the tasks involved in intricate biological processes. To cope with this challenge, we propose using high-content cell microscopy images to assist in learning molecular representation. The fundamental rationale of our method is to leverage the correspondence between molecular topological structures and the caused perturbations at the phenotypic level. By including cross-modal pre-training with different types of contrastive loss functions in a unified framework, our model can efficiently learn generic and informative representations from cellular images, which are complementary to molecular structures. Empirical experiments demonstrated that the model transfers non-trivially to a variety of downstream tasks and is often competitive with the existing SSL baselines, e.g., a 15.4\% absolute Hit@10 gains in graph-image retrieval task and a 4.0\% absolute AUC improvements in clinical outcome predictions. Further zero-shot case studies show the potential of the approach to be applied to real-world drug discovery.
****

A Closer Look at Self-supervised Lightweight Vision Transformers
Shaoru Wang,Jin Gao,Zeming Li,Weiming Hu
Self-supervised learning on large-scale Vision Transformers (ViTs) as pre-training methods has achieved promising downstream performance. Yet, how much these pre-training paradigms promote lightweight ViTs' performance is considerably less studied. In this work, we mainly develop and benchmark self-supervised pre-training methods, e.g., contrastive-learning-based MoCo-v3, masked-image-modeling-based MAE on image classification tasks, and some downstream dense prediction tasks. We surprisingly find that if proper pre-training is adopted, even vanilla lightweight ViTs show comparable performance on ImageNet to previous SOTA networks with delicate architecture design. We also point out some defects of such pre-training, \eg, failing to benefit from large-scale pre-training data and showing inferior performance on data-insufficient downstream tasks. Furthermore, we analyze and clearly show the effect of such pre-training by analyzing the properties of the layer representation and attention maps for related models. Finally, based on the above analyses, a distillation strategy during pre-training is developed, which leads to further downstream performance improvement for MAE-based pre-training.
****

MANDERA: Malicious Node Detection in Federated Learning via Ranking
Wanchuang Zhu,Benjamin Zi Hao Zhao,Simon Luo,Tongliang Liu,Ke Deng
Byzantine attacks hinder the deployment of federated learning algorithms. Although we know that the benign gradients and Byzantine attacked gradients are distributed differently, to detect the malicious gradients is challenging due to (1) the gradient is high-dimensional and each dimension has its unique distribution and (2) the benign gradients and the attacked gradients are always mixed (two-sample test methods cannot apply directly). To address the above, for the first tim

e, we propose MANDERA which is theoretically guaranteed to efficiently detect all malicious gradients under Byzantine attacks with no prior knowledge or history about the number of attacked nodes.
Specifically, we transfer the original updating gradient matrix into a ranking matrix. By such an operation, the scales of different dimensions of the gradients in the ranking space become identical. The high-dimensional benign gradients and the malicious gradients can be easily separated. The effectiveness of MANDERA is further confirmed by experimentation on four Byzantine attack implementations (Gaussian, Zero Gradient, Sign Flipping, Shifted Mean), comparing with state-of-the-art defenses. The experiments cover both IID and Non-IID datasets.

****

MQSP: Micro-Query Sequence Parallelism for Linearly Scaling Long Sequence Transformer

Ying Zhong,Jianjiang Zhu,PengCheng Yang,Xiaoming Zhang,Ke Zhang

Long sequence modeling of Transformer gains prevalence in fields involving long texts and high-resolution images and videos but suffers from quadratic memory complexity. Existing work investigates low-complexity variants or parallel methods to handle it. The former attempts to approximate full attention and is limited by a single device's capacity. The latter struggles to manage quadratic memory of attention maps, leading to insufficient sequence scalability. In this work, we propose a novel parallel method named $\textbf{M}$icro-$\textbf{Q}$uery $\textbf{S}$equence $\textbf{P}$arallelism. MQSP slices sequences across devices and projects local queries, keys, and values in self-attention. For communication and memory efficiency, MQSP all-gathers the queries while keys and values remain locally to acquire the local attention map, on which a distributed softmax gets conducted to amortize memory by column. Meanwhile, the queries get further partitioned as Micro-Q to divide the computation and recycle the attention map by row, jointly decomposing the quadratic memory to achieve linear scalability. The evaluation result shows that MQSP scales up sequence length linearly and achieves 4.5$\times$ sequence length of ColossalAI's sequence parallelism and 4.3$\times$ of Megatron-LM3, enabling training BERT-large of 78848 sequence length on 32 A100 GPUs. MQSP can reduce up to 78.6$\%$ of memory occupation and achieve up to 3.3$\times$ throughput when training on 17408 sequence length. The convergence quality experiment proves that MQSP provides the means for long sequences with guaranteed convergence, bringing the potential for the Transformer to explore longer sequences.

****

DSPNet: Towards Slimmable Pretrained Networks based on Discriminative Self-supervised Learning

Shaoru Wang,Zeming Li,Jin Gao,Weiming Hu

Self-supervised learning (SSL) has achieved promising downstream performance. However, when facing various resource budgets in real-world applications, it costs a huge computation burden to pretrain multiple networks of various sizes one by one. In this paper, we propose Discriminative-SSL-based Slimmable Pretrained Networks (DSPNet), which can be trained once and then slimmed to multiple sub-networks of various sizes, each of which faithfully learns good representation and can serve as good initialization for downstream tasks with various resource budgets. Specifically, we extend the idea of slimmable networks to a discriminative SSL paradigm, by integrating SSL and knowledge distillation gracefully. We show comparable or improved performance of DSPNet on ImageNet to the networks individually pretrained one by one under the linear evaluation and semi-supervised evaluation protocols, while reducing large training cost. The pretrained models also generalize well on downstream detection and segmentation tasks. Code will be made public.

****

Generative Multi-Flow Networks: Centralized, Independent and Conservation

Yinchuan Li,Haozhi Wang,Shuang Luo,yunfeng shao,Jianye HAO

Generative flow networks utilize the flow matching loss to learn a stochastic policy for generating objects from a sequence of actions, such that the probability of generating a pattern can be proportional to the corresponding given reward.

However, existing works can only handle single flow model tasks and cannot directly generalize to multi-agent flow networks due to limitations such as flow estimation complexity and independent sampling. In this paper, we propose the framework of generative multi-flow networks (GMFlowNets) that can be applied to multiple agents to generate objects collaboratively through a series of joint actions. Then, the centralized flow network algorithm is proposed for centralized training GMFlowNets, while the independent flow network algorithm is proposed to achieve decentralized execution of GMFlowNets. Based on the independent global conservation condition, the flow conservation network algorithm is then proposed to realize centralized training with decentralized execution paradigm. Theoretical analysis proves that using the multi-flow matching loss function can train a unique Markovian flow, and the flow conservation network can ensure independent policies can generate samples with probability proportional to the reward function. Experimental results demonstrate the performance superiority of the proposed algorithms compared to reinforcement learning and MCMC-based methods.
****

A Laplace-inspired Distribution on SO(3) for Probabilistic Rotation Estimation
Yingda Yin,Yang Wang,He Wang,Baoquan Chen
Estimating the 3DoF rotation from a single RGB image is an important yet challenging problem. Probabilistic rotation regression has raised more and more attention with the benefit of expressing uncertainty information along with the prediction. Though modeling noise using Gaussian-resembling Bingham distribution and matrix Fisher distribution is natural, they are shown to be sensitive to outliers for the nature of quadratic punishment to deviations. In this paper, we draw inspiration from multivariate Laplace distribution and propose a novel Rotation Laplace distribution on SO(3). Rotation Laplace distribution is robust to the disturbance of outliers and enforces much gradient to the low-error region, resulting in a better convergence. Our extensive experiments show that our proposed distribution achieves state-of-the-art performance for rotation regression tasks over both probabilistic and non-probabilistic baselines. Our project page is at pku-epic.github.io/RotationLaplace.
****

ContraGen: Effective Contrastive Learning For Causal Language Model
Nihal Jain,Dejiao Zhang,Wasi Uddin Ahmad,Zijian Wang,Feng Nan,Xiaopeng Li,Ming Tan,Ramesh Nallapati,Baishakhi Ray,Parminder Bhatia,Xiaofei Ma,Bing Xiang
Despite exciting progress in large-scale language generation, the expressiveness of its representations is severely limited by the \textit{anisotropy} issue where the hidden representations are distributed into a narrow cone in the vector space. To address this issue, we present ContraGen, a novel contrastive learning framework to improve the representation with better uniformity and discrimination at both sequence-level and token-level. We assess ContraGen on a wide range of downstream tasks in natural and programming languages. We show that ContraGen can effectively enhance both uniformity and discrimination of the representations and lead to the desired improvement on various language understanding tasks where discriminative representations are crucial for attaining good performance. Specifically, we attain 45.9% relative improvement on the Semantic Textual Similarity tasks and 33.5% on Code-to-Code Search tasks. Furthermore, by improving the expressiveness of the representations, ContraGen also boosts the source code generation capability with 9% relative improvement on execution accuracy on HumanEval benchmark.
****

Time Series Anomaly Detection via Hypothesis Testing for Dynamical Systems
Haowei He,Jingzhao Zhang,Yanan Wang,Benben Jiang,Shaobo Huang,Chen Wang,Yang Zhang,Xuebing Han,Dongxu Guo,Guannan He,Minggao Ouyang
Real world systems---such as robots, weather, energy systems and stock markets---are complicated and high-dimensional. Hence, without prior knowledge of the system dynamics, detecting or forecasting abnormal events from the sequential observations of the system is challenging. In this work, we address the problem caused by high-dimensionality via viewing time series anomaly detection as hypothesis testing on dynamical systems. This perspective can avoid the dimension of the

problem from increasing linearly with time horizon, and naturally leads to a novel anomaly detection model, termed as DyAD (Dynamical system Anomaly Detection). Furthermore, as existing time-series anomaly detection algorithms are usually evaluated on relatively small datasets, we released a large-scale one on detecting battery failures in electric vehicles. We benchmarked several popular algorithms on both public datasets and our released new dataset. Our experiments demonstrated that our proposed model achieves state-of-the-art results.

****

Schrödinger's FP: Training Neural Networks with Dynamic Floating-Point Containers

Milos Nikolic,Enrique Torres,Jiahui Wang,Ali Hadi Zadeh,Mostafa Mahmoud,Ameer Abdelhadi,Andreas Moshovos

We introduce a software-hardware co-design approach to reduce memory traffic and footprint during training with BFloat16 or FP32, in order to boost energy efficiency and execution time performance. Our methods dynamically adjust the size and format of the floating-point containers used to store activations and weights during training. The different value distributions lead us to different approaches for exponents and mantissas. Gecko exploits the favourable exponent distribution with a lossless delta encoding approach to reduce the total exponent footprint by up to 58% in comparison to the FP32 baseline. To contend with the noisy mantissa distributions, we present two lossy methods to eliminate as many as possible least significant bits without affecting accuracy. Quantum Mantissa is a machine learning mantissa compression method that taps onto the gradient descent algorithm to learn the minimal mantissa bitlengths on a per-layer granularity, and obtain up to 92% reduction in total mantissa footprint. Alternatively, BitChop observes changes in the loss function during training to adjust mantissa bitlength network-wide, yielding a reduction of 81% in footprint. Schrödinger's FP implements hardware encoders/decoders that, guided by Gecko/Quantum Mantissa or Gecko/BitChop, transparently encode/decode values when transferring to/from off-chip memory, boosting energy efficiency and reducing execution time.

****

Measuring and Narrowing the Compositionality Gap in Language Models

Ofir Press,Muru Zhang,Sewon Min,Ludwig Schmidt,Noah A. Smith,Mike Lewis

We investigate the ability of language models to perform compositional reasoning tasks where the overall solution depends on correctly composing the answers to sub-problems. We measure how often models can correctly answer all sub-problems but not generate the overall solution, a ratio we call the compositionality gap. We evaluate this ratio by asking multi-hop questions with answers that require composing multiple facts unlikely to have been observed together during pretraining. In the GPT-3 family of models, as model size increases we show that the single-hop question answering performance improves faster than the multi-hop performance does, therefore the compositionality gap does not decrease. This surprising result suggests that while more powerful models memorize and recall more factual knowledge, they show no corresponding improvement in their ability to perform this kind of compositional reasoning.

We then demonstrate how elicitive prompting (such as chain of thought) narrows the compositionality gap by reasoning explicitly instead of implicitly. We present a new method, self-ask, that further improves on chain of thought. In our method, the model explicitly asks itself (and then answers) follow-up questions before answering the initial question. We finally show that self-ask's structured prompting lets us easily plug in a search engine to answer the follow-up questions, which additionally improves accuracy.

****

HiViT: A Simpler and More Efficient Design of Hierarchical Vision Transformer

Xiaosong Zhang,Yunjie Tian,Lingxi Xie,Wei Huang,Qi Dai,Qixiang Ye,Qi Tian

There has been a debate on the choice of plain vs. hierarchical vision transformers, where researchers often believe that the former (e.g., ViT) has a simpler design but the latter (e.g., Swin) enjoys higher recognition accuracy. Recently, the emerge of masked image modeling (MIM), a self-supervised visual pre-training method, raised a new challenge to vision transformers in terms of flexibility, i.e., part of image patches or tokens are to be discarded, which seems to claim the advantages of plain vision transformers. In this paper, we delve deep into the comparison between ViT and Swin, revealing that (i) the performance gain of Swin is mainly brought by a deepened backbone and relative positional encoding, (ii) the hierarchical design of Swin can be simplified into hierarchical patch embedding (proposed in this work), and (iii) other designs such as shifted-window attentions can be removed. By removing the unnecessary operations, we come up with a new architecture named HiViT (short for hierarchical ViT), which is simpler and more efficient than Swin yet further improves its performance on fully-supervised and self-supervised visual representation learning. In particular, after pre-trained using masked autoencoder (MAE) on ImageNet-1K, HiViT-B reports a 84.6% accuracy on ImageNet-1K classification, a 53.3% box AP on COCO detection, and a 52.8% mIoU on ADE20K segmentation, significantly surpassing the baseline. Code is available at https://github.com/zhangxiaosong18/hivit.

****

Style Spectroscope: Improve Interpretability and Controllability through Fourier Analysis

Zhiyu Jin,Xuli Shen,Bin Li,Xiangyang Xue

Universal style transfer (UST) infuses styles from arbitrary reference images into content images. Existing methods, while enjoying many practical successes, are unable of explaining experimental observations, including different performances of UST algorithms in preserving the spatial structure of content images. In addition, methods are limited to cumbersome global controls on stylization, so that they require additional spatial masks for desired stylization. In this work, we provide a systematic Fourier analysis on a general framework for UST. We present an equivalent form of the framework in the frequency domain. The form implies that existing algorithms treat all frequency components and pixels of feature maps equally, except for the zero-frequency component. We connect Fourier amplitude and phase with Gram matrices and a content reconstruction loss in style transfer, respectively. Based on such equivalence and connections, we can thus interpret different structure preservation behaviors between algorithms with Fourier phase. Given the interpretations we have, we propose two manipulations in practice for structure preservation and desired stylization. Both qualitative and quantitative experiments demonstrate the competitive performance of our method against the state-of-the-art methods. We also conduct experiments to demonstrate (1) the abovementioned equivalence, (2) the interpretability based on Fourier amplitude and phase and (3) the controllability associated with frequency components.

****

Multimodal Federated Learning via Contrastive Representation Ensemble

Qiying Yu,Yang Liu,Yimu Wang,Ke Xu,Jingjing Liu

With the increasing amount of multimedia data on modern mobile systems and IoT infrastructures, harnessing these rich multimodal data without breaching user privacy becomes a critical issue. Federated learning (FL) serves as a privacy-conscious alternative to centralized machine learning. However, existing FL methods extended to multimodal data all rely on model aggregation on single modality level, which restrains the server and clients to have identical model architecture for each modality. This limits the global model in terms of both model complexity and data capacity, not to mention task diversity. In this work, we propose \textit{Contrastive Representation Ensemble and Aggregation for Multimodal FL (CreamFL)}, a multimodal federated learning framework that enables training larger server models from clients with heterogeneous model architectures and data modalities, while only communicating knowledge on public dataset. To achieve better multimodal representation fusion, we design a global-local cross-modal ensemble strategy to aggregate client representations. To mitigate local model drift caused by two unprecedented heterogeneous factors stemming from multimodal discrepancy (\textit{modality gap} and \textit{task gap}), we further propose two inter-modal and intra-modal contrasts to regularize local training, which complements information of the absent modality for uni-modal clients and regularizes local clients to head towards global consensus. Thorough evaluations and ablation studies on image-text retrieval and visual question answering tasks showcase the superiori ty of CreamFL over state-of-the-art FL methods and its practical value.

****

Eva: Practical Second-order Optimization with Kronecker-vectorized Approximation
Lin Zhang,Shaohuai Shi,Bo Li

Second-order optimization algorithms exhibit excellent convergence properties for training deep learning models, but often incur significant computation and memory overheads. This can result in lower training efficiency than the first-order counterparts such as stochastic gradient descent (SGD). In this work, we present a memory- and time-efficient second-order algorithm named Eva with two novel techniques: 1) we construct the second-order information with the Kronecker factorization of small stochastic vectors over a mini-batch of training data to reduce memory consumption, and 2) we derive an efficient update formula without explicitly computing the inverse of matrices using the Sherman-Morrison formula. We further provide a theoretical interpretation of Eva from a trust-region optimization point of view to understand how it works. Extensive experimental results on different models and datasets show that Eva reduces the end-to-end training time up to $2.05\times$ and $2.42\times$ compared to first-order SGD and second-order algorithms (K-FAC and Shampoo), respectively.

****

Identifying Weight-Variant Latent Causal Models
Yuhang Liu,Zhen Zhang,Dong Gong,Mingming Gong,Biwei Huang,Anton van den Hengel,Kun Zhang,Javen Qinfeng Shi

The task of causal representation learning aims to uncover latent higher-level causal representations that affect lower-level observations. Identifying true latent causal representations from observed data, while allowing instantaneous causal relations among latent variables, remains a challenge, however. To this end, we start from the analysis of three intrinsic properties in identifying latent space from observations: transitivity, permutation indeterminacy, and scaling indeterminacy. We find that transitivity acts as a key role in impeding the identifiability of latent causal representations. To address the unidentifiable issue due to transitivity, we introduce a novel identifiability condition where the underlying latent causal model satisfies a linear-Gaussian model, in which the causal coefficients and the distribution of Gaussian noise are modulated by an additional observed variable. Under some mild assumptions, we can show that the latent causal representations can be identified up to trivial permutation and scaling. Furthermore, based on this theoretical result, we propose a novel method, termed Structural caUsAl Variational autoEncoder, which directly learns latent causal representations and causal relationships among them, together with the mapping from the latent causal variables to the observed ones. We show that the proposed method learns the true parameters asymptotically. Experimental results on synthetic and real data demonstrate the identifiability and consistency results and the efficacy of the proposed method in learning latent causal representations.

****

Beyond Single Path Integrated Gradients for Reliable Input Attribution via Randomized Path Sampling
Giyoung Jeon,Haedong Jeong,Jaesik Choi

Input attribution is a widely used explanation method for deep neural networks, especially in visual tasks. Among various attribution methods, Integrated Gradients (IG) is frequently used because of its model-agnostic applicability and desirable axioms. However, previous work has shown that such method often produces noisy and unreliable attributions during the integration of the gradients over the path defined in the input space. In this paper, we tackle this issue by estimating the distribution of the possible attributions according to the integrating path selection. We show that such noisy attribution can be reduced by aggregating attributions from the multiple paths instead of using a single path. Inspired by Stick-Breaking Process, we suggest a random process to generate rich and various sampling of the gradient integrating path. Using multiple input attributions obtained from randomized path, we propose a novel attribution measure using the distribution of attributions at each input features. We identify proposed method qualitatively show less-noisy and object-aligned attribution and its feasibili

ty through the quantitative evaluations.
****

Sweet Gradient Matters: Designing Consistent and Efficient Estimator for Zero-Shot Neural Architecture Search

Longxing Yang,Yanxin Fu,Shun Lu,Zihao Sun,Jilin Mei,Wenxiao Zhao,Yu Hu

Neural architecture search (NAS) is one of the core technologies of AutoML for designing high-performance networks. Recently, Zero-Shot NAS has gained growing interest due to its training-free property and super-fast search speed. However, existing Zero-Shot estimators commonly suffer from low consistency, which limits the reliability and applicability. In this paper, we observe that Sweet Gradient of parameters, i.e., the absolute gradient values within a certain interval, brings higher consistency in network performance compared to the overall number of parameters. We further demonstrate a positive correlation between the network depth and the parameter ratio of sweet gradients in each layer. Based on the analysis, we propose a training-free method to find the Sweet Gradient interval and obtain an estimator, named Sweetimator. Experiments show that Sweetimator has superior consistency to existing Zero-Shot estimators on four benchmarks with eight search spaces. Moreover, Sweetimator achieves state-of-the-art performance on NAS-Bench-201 and DARTS search spaces.
****

Neural Collaborative Filtering Bandits via Meta Learning

Yikun Ban,Yunzhe Qi,Tianxin Wei,Jingrui He

Contextual multi-armed bandits provide powerful tools to solve the exploitation-exploration dilemma in decision making, with direct applications in the personalized recommendation. In fact, collaborative effects among users carry the significant potential to improve the recommendation. In this paper, we introduce and study the problem by exploring `Neural Collaborative Filtering Bandits', where the rewards can be non-linear functions and groups are formed dynamically given different specific contents. To solve this problem, we propose a meta-learning based bandit algorithm, Meta-Ban (\textbf{meta-ban}dits), where a meta-learner is designed to represent and rapidly adapt to dynamic groups, along with an informative UCB-based exploration strategy. Furthermore, we analyze that Meta-Ban can achieve the regret bound of $\mathcal{O}(\sqrt{nT\log T})$, which is sharper over state-of-the-art related works. In the end, we conduct extensive experiments showing that Meta-Ban outperforms six strong baselines.
****

Cascaded Teaching Transformers with Data Reweighting for Long Sequence Time-series Forecasting

Haoyi Zhou,Chonghan Gao,Pengtao Xie,Jianxin Li

The Transformer-based models have shown superior performance in the long sequence time-series forecasting problem. The sparsity assumption on self-attention dot-product reveals that not all inputs are equally significant for Transformers. Instead of implicitly utilizing weighted time-series, we build a new learning framework by cascaded teaching Transformers to reweight samples. We formulate the framework as a multi-level optimization and design three different dataset-weight generators. We perform extensive experiments on five datasets, which shows that our proposed method could significantly outperform the SOTA Transformers.
****

Can CNNs Be More Robust Than Transformers?

Zeyu Wang,Yutong Bai,Yuyin Zhou,Cihang Xie

The recent success of Vision Transformers is shaking the long dominance of Convolutional Neural Networks (CNNs) in image recognition for a decade. Specifically, in terms of robustness on out-of-distribution samples, recent research finds that Transformers are inherently more robust than CNNs, regardless of different training setups. Moreover, it is believed that such superiority of Transformers should largely be credited to their \emph{self-attention-like architectures per se}. In this paper, we question that belief by closely examining the design of Transformers. Our findings lead to three highly effective architecture designs for boosting robustness, yet simple enough to be implemented in several lines of code, namely a) patchifying input images, b) enlarging kernel size, and c) reducing activation layers and normalization layers. Bringing these components together, we are able to build pure CNN architectures without any attention-like operations that are as robust as, or even more robust than, Transformers. We hope this work can help the community better understand the design of robust neural architectures. The code is publicly available at https://github.com/UCSC-VLAA/RobustCNN.

****

Decoupled Mixup for Data-efficient Learning

Zicheng Liu,Siyuan Li,Ge Wang,Cheng Tan,Lirong Wu,Stan Z. Li

Mixup is an efficient data augmentation approach that improves the generalization of neural networks by smoothing the decision boundary with mixed data. Recently, dynamic mixup methods have improved previous static policies effectively (e.g., linear interpolation) by maximizing salient regions or maintaining the target in mixed samples. The discrepancy is that the generated mixed samples from dynamic policies are more instance discriminative than the static ones, e.g., the foreground objects are decoupled from the background. However, optimizing mixup policies with dynamic methods in input space is an expensive computation compared to static ones. Hence, we are trying to transfer the decoupling mechanism of dynamic methods from the data level to the objective function level and propose the general decoupled mixup (DM) loss. The primary effect is that DM can adaptively focus on discriminative features without losing the original smoothness of the mixup while avoiding heavy computational overhead. As a result, DM enables static mixup methods to achieve comparable or even exceed the performance of dynamic methods. This also leads to an interesting objective design problem for mixup training that we need to focus on both smoothing the decision boundaries and identifying discriminative features. Extensive experiments on supervised and semi-supervised learning benchmarks across seven classification datasets validate the effectiveness of DM by equipping it with various mixup methods.

****

FAIRER: Fairness as Decision Rationale Alignment

Tianlin Li,Qing Guo

Deep neural networks (DNNs) have achieved remarkable accuracy, but they often suffer from fairness issues, as deep models typically show distinct accuracy differences among some specific subgroups (e.g., males and females). Existing research addresses this critical issue by employing fairness-aware loss functions to constrain the last-layer outputs and directly regularize DNNs. Although the fairness of DNNs is improved, it is unclear how the trained network makes a fair prediction, which limits future fairness improvements. In this paper, we investigate fairness from the perspective of decision rationale and define neuron parity scores to characterize the fair decision process of networks by analyzing neuron behaviors in various subgroups. Extensive empirical studies show that the unfair issue could arise from the unaligned decision rationales of subgroups. Existing fairness regularization terms fail to achieve decision rationale alignment because they only constrain last-layer outputs while ignoring intermediate neuron alignment. To address the issue, we formulate the fairness as a new task, i.e., decision rationale alignment that requires DNNs’ neurons to have consistent responses on subgroups at both intermediate processes and the final prediction. To make this idea practical during optimization, we relax the naive objective function and propose gradient-guided parity alignment, which encourages gradient-weighted consistency of neurons across subgroups. Extensive experiments on a variety of datasets show that our method can improve fairness while maintaining high accuracy and outperforming other baselines by a large margin. We have released our codes at https://anonymous.4open.science/r/fairer_submission-F176/.

****

Risk-Aware Reinforcement Learning with Coherent Risk Measures and Non-linear Function Approximation

Thanh Lam,Arun Verma,Bryan Kian Hsiang Low,Patrick Jaillet

We study the risk-aware reinforcement learning (RL) problem in the episodic finite-horizon Markov decision process with unknown transition and reward functions. In contrast to the risk-neutral RL problem, we consider minimizing the risk of having low rewards, which arise due to the intrinsic randomness of the MDPs and imperfect knowledge of the model. Our work provides a unified framework to analyze the regret of risk-aware RL policy with coherent risk measures in conjunction with non-linear function approximation, which gives the first sub-linear regret bounds in the setting. Finally, we validate our theoretical results via empirical experiments on synthetic and real-world data.

****

A Minimalist Dataset for Systematic Generalization of Perception, Syntax, and Semantics

Qing Li,Siyuan Huang,Yining Hong,Yixin Zhu,Ying Nian Wu,Song-Chun Zhu

Inspired by humans' exceptional ability to master arithmetic and generalize to new problems, we present a new dataset, HINT, to examine machines' capability of learning generalizable concepts at three levels: perception, syntax, and semantics. In HINT, machines are tasked with learning how concepts are perceived from raw signals such as images (i.e., perception), how multiple concepts are structurally combined to form a valid expression (i.e., syntax), and how concepts are realized to afford various reasoning tasks (i.e., semantics), all in a weakly supervised manner. Focusing on systematic generalization, we carefully design a five-fold test set to evaluate both the interpolation and the extrapolation of learned concepts w.r.t the three levels. Further, we design a few-shot learning split to determine whether or not models can rapidly learn new concepts and generalize them to more complex scenarios. To comprehend existing models' limitations, we undertake extensive experiments with various sequence-to-sequence models, including RNNs, Transformers, and GPT-3 (with the chain of thought prompting). The results indicate that current models struggle to extrapolate to long-range syntactic dependency and semantics. Models exhibit a considerable gap toward human-level generalization when evaluated with new concepts in a few-shot setting. Moreover, we discover that it is infeasible to solve HINT by merely scaling up the dataset and the model size; this strategy contributes little to the extrapolation of syntax and semantics. Finally, in zero-shot GPT-3 experiments, the chain of thought prompting exhibits impressive results and significantly boosts the test accuracy. We believe the HINT dataset and the experimental findings are of great interest to the learning community on systematic generalization.%

****

Bi-level Physics-Informed Neural Networks for PDE Constrained Optimization using Broyden's Hypergradients

Zhongkai Hao,Chengyang Ying,Hang Su,Jun Zhu,Jian Song,Ze Cheng

Deep learning based approaches like Physics-informed neural networks (PINNs) and DeepONets have shown promise on solving PDE constrained optimization (PDECO) problems.

However, existing methods are insufficient to handle those PDE constraints that have a complicated or nonlinear dependency on optimization targets. In this paper, we present a novel bi-level optimization framework to resolve the challenge by decoupling the optimization of the targets and constraints. For the inner loop optimization, we adopt PINNs to solve the PDE constraints only. For the outer loop, we design a novel method by using Broyden's method based on the Implicit Function Theorem (IFT), which is efficient and accurate for approximating hypergradients. We further present theoretical explanations and error analysis of the hypergradients computation. Extensive experiments on multiple large-scale and nonlinear PDE constrained optimization problems demonstrate that our method achieves state-of-the-art results compared with strong baselines.

****

Hazard Gradient Penalty for Survival Analysis

Seungjae Jung,Kyung-Min Kim

Survival analysis appears in various fields such as medicine, economics, engineering, and business.

Recent studies showed that the Ordinary Differential Equation (ODE) modeling framework integrates many existing survival models while the framework is flexible and widely applicable.

However, naively applying the ODE framework to survival analysis problems may mo

del fiercely changing density function with respect to covariates which may worsen the model’s performance.
Though we can apply L1 or L2 regularizers to the ODE model, their effect on the ODE modeling framework is barely known.
In this paper, we propose hazard gradient penalty (HGP) to enhance the performance of a survival analysis model.
Our method imposes constraints on local data points by regularizing the gradient of hazard function with respect to the data point.
Our method applies to any survival analysis model including the ODE modeling framework and is easy to implement.
We theoretically show that our method is related to minimizing the KL divergence between the density function at a data point and that of the neighborhood points.
Experimental results on three public benchmarks show that our approach outperforms other regularization methods.
****
Rethink Depth Separation with Intra-layer Links
FENGLEI FAN,Zeyu Li,Huan Xiong,Tieyong Zeng
The depth separation theory is nowadays widely accepted as an effective explanation for the power of depth, which consists of two parts: i) there exists a function representable by a deep network; ii) such a function cannot be represented by a shallow network whose width is lower than a threshold. Here, we report that adding intra-layer links can greatly improve a network's representation capability through the bound estimation, explicit construction, and functional space analysis. Then, we modify the depth separation theory by showing that a shallow network with intra-layer links does not need to go as wide as before to express some hard functions constructed by a deep network. Such functions include the renowned "sawtooth" functions. Our results supplement the existing depth separation theory by examining its limit in a broader domain. Also, our results suggest that once configured with an appropriate structure, a shallow and wide network may have expressive power on a par with a deep network.
****
Reach the Remote Neighbors: Dual-Encoding Transformer for Graphs
Lingbing Guo,Zequn Sun,Qiang Zhang,Huajun Chen
Despite recent successes in natural language processing and computer vision, Transformer suffers from the scalability problem when dealing with graphs. Computing node-to-node attentions is infeasible on complicated graphs, e.g., knowledge graphs. One solution is to consider only the near neighbors, which, however, will lose the key merit of Transformer that attends to the elements at any distance. In this paper, we propose a new Transformer architecture, named dual-encoding Transformer (DET), which has a structural encoder to aggregate information from near neighbors and a semantic encoder to focus on useful semantically close neighbors. The two encoders can be incorporated to boost each other's performance. Our experiments demonstrate that DET achieves superior performance compared to the respective state-of-the-art attention-based methods in dealing with molecules, networks and knowledge graphs.
****
The Geometry of Self-supervised Learning Models and its Impact on Transfer Learning
Romain Cosentino,Sarath Shekkizhar,Mahdi Soltanolkotabi,Salman Avestimehr,Antonio Ortega
Self-supervised learning~(SSL) has emerged as a desirable paradigm in computer vision due to the inability of supervised models to learn representations that can generalize in domains with limited labels. The recent popularity of SSL has led to the development of several models that make use of diverse training strategies, architectures, and data augmentation policies with no existing unified framework to study or assess their effectiveness in transfer learning.
We propose a data-driven geometric strategy to analyze different SSL models using local neighborhoods in the feature space induced by each. Unlike existing approaches that consider mathematical approximations of the parameters, individual c

omponents, or optimization landscape, our work aims to explore the geometric properties of the representation manifolds learned by SSL models.
Our proposed manifold graph metrics~(MGMs)
provide insights into the geometric similarities and differences between available SSL models, their invariances with respect to specific augmentations, and their performances on transfer learning tasks. Our key findings are two fold: (i) contrary to popular belief, the geometry of SSL models is not tied to its training paradigm (contrastive, non-contrastive, and cluster-based); (ii) we can predict the transfer learning capability for a specific model based on the geometric properties of its semantic and augmentation manifolds.
****

When Do Models Generalize? A Perspective From Data-Algorithm Compatibility
Jing Xu,Jiaye Teng,Yang Yuan,Andrew C Yao
One of the major open problems in machine learning is to characterize generalization in the overparameterized regime, where most traditional generalization bounds become inconsistent (Nagarajan and Kolter, 2019). In many scenarios, their failure can be attributed to obscuring the crucial interplay between the training algorithm and the underlying data distribution. To address this issue, we propose a concept named compatibility, which quantitatively characterizes generalization in a both data-relevant and algorithm relevant manner. By considering the entire training trajectory and focusing on early-stopping iterates, compatibility exploits the data and the algorithm information and is therefore a more suitable notion for generalization. We validate this by theoretically studying compatibility under the setting of solving overparameterized linear regression with gradient descent. Specifically, we perform a data-dependent trajectory analysis and derive a sufficient condition for compatibility in such a setting. Our theoretical results demonstrate that in the sense of compatibility, generalization holds with significantly weaker restrictions on the problem instance than the previous last iterate analysis.
****

On the Saturation Effect of Kernel Ridge Regression
Yicheng Li,Haobo Zhang,Qian Lin
The saturation effect refers to the phenomenon that the kernel ridge regression (KRR) fails to achieve the information theoretical lower bound when the smoothness of the underground truth function exceeds certain level. The saturation effect has been widely observed in practices and a saturation lower bound of KRR has been conjectured for decades. In this paper, we provide a proof of this long-standing conjecture.
****

Adversarial perturbation based latent reconstruction for domain-agnostic self-supervised learning
Sijie Tian,Kuilin Chen,Chi-Guhn Lee
Most self-supervised learning (SSL) methods rely on domain-specific pretext tasks and data augmentations to learn high-quality representations from unlabeled data. Development of those pretext tasks and data augmentations requires expert domain knowledge. In addition, it is not clear why solving certain pretext tasks leads to useful representations. Those two reasons hinder wider application of SSL to different domains. To overcome such limitations, we propose adversarial perturbation based latent reconstruction (APLR) for domain-agnostic self-supervised learning. In APLR, a neural network is trained to generate adversarial noise to perturb the unlabeled training sample so that domain-specific augmentations are not required. The pretext task in APLR is to reconstruct the latent representation of a clean sample from a perturbed sample. We show that representation learning via latent reconstruction is closely related to multi-dimensional Hirschfeld-Gebelein-Rényi (HGR) maximal correlation and has theoretical guarantees on the linear probe error. To demonstrate the effectiveness of APLR, the proposed method is applied to various domains such as tabular data, images, and audios. Empirical results indicate that APLR not only outperforms existing domain-agnostic SSL methods, but also closes the performance gap to domain-specific SSL methods. In many cases, APLR also outperforms training the full network in a supervised man

ner.
****

Unsupervised Model Selection for Time Series Anomaly Detection

Mononito Goswami,Cristian Ignacio Challu,Laurent Callot,Lenon Minorics,Andrey Kan

Anomaly detection in time-series has a wide range of practical applications. While numerous anomaly detection methods have been proposed in the literature, a recent survey concluded that no single method is the most accurate across various datasets. To make matters worse, anomaly labels are scarce and rarely available in practice. The practical problem of selecting the most accurate model for a given dataset without labels has received little attention in the literature. This paper answers this question \textit{i.e.} Given an unlabeled dataset and a set of candidate anomaly detectors, how can we select the most accurate model? To this end, we identify three classes of surrogate (unsupervised) metrics, namely, \textit{prediction error}, \textit{model centrality}, and \textit{performance on injected synthetic anomalies}, and show that some metrics are highly correlated with standard supervised anomaly detection performance metrics such as the F_1 score, but to varying degrees. We formulate metric combination with multiple imperfect surrogate metrics as a robust rank aggregation problem. We then provide theoretical justification behind the proposed approach. Large-scale experiments on multiple real-world datasets demonstrate that our proposed unsupervised approach is as effective as selecting the most accurate model based on partially labeled data.

****

Constrained Hierarchical Deep Reinforcement Learning with Differentiable Formal Specifications

Zikang Xiong,Joe Eappen,Ahmed H Qureshi,Suresh Jagannathan

Formal logic specifications are a useful tool to describe desired agent behavior and have been explored as a means to shape rewards in Deep Reinforcement Learning (DRL) systems over a variety of problems and domains. Prior work, however, has failed to consider the possibility of making these specifications differentiable, which would yield a more informative signal of the objective via the specification gradient. This paper examines precisely such an approach by exploring a Lagrangian method to constrain policy updates using a differentiable style of temporal logic specifications that associates logic formulae with real-valued quantitative semantics. This constrained learning mechanism is then used in a hierarchical setting where a high-level specification-guided neural network path planner works with a low-level control policy to navigate through planned waypoints. The effectiveness of our approach is demonstrated over four robot dynamics with five different types of Linear Temporal Logic (LTL) specifications. Our demo videos are collected at https://sites.google.com/view/schrl.

****

Topic Aware Transformer: Domain Shift for Unconditional Text Generation Model

Noriaki Kawamae

Our goal is to adapt pre-trained language models (PLMs) to support unconditional text generation tasks.
Because Transformer-based models are pre-trained on more massive and heterogeneous corpora than specific target corpus,
the gap between these corpora and the target corpus raises the question of whether these PLMs will actually benefit this task even after fine-tuning.
As the domain adaptation of PLMs needs to bridge this gap,
we propose a framework, Topic Aware Transformer (TAT), that adapts PLMs for target-aware text generation while alleviating catastrophic forgetting.
The motivation of TAT to distill the target-specific knowledge as topics,
and steer PLMs toward these topics.
This requirement and motivation lead us to introduce a topic steering layer (TSL) as an additional layer,
and Topic Distribution Modeling (TDM) as a training task.
Experiments show that these components resolve the gap as the domain shift,
and can tailor PLMs to generate text to better reflect a given small fine-tuning

corpus.
****

Protein Representation Learning by Geometric Structure Pretraining
Zuobai Zhang,Minghao Xu,Arian Rokkum Jamasb,Vijil Chenthamarakshan,Aurelie Lozano,Payel Das,Jian Tang

Learning effective protein representations is critical in a variety of tasks in biology such as predicting protein function or structure. Existing approaches usually pretrain protein language models on a large number of unlabeled amino acid sequences and then finetune the models with some labeled data in downstream tasks. Despite the effectiveness of sequence-based approaches, the power of pretraining on known protein structures, which are available in smaller numbers only, has not been explored for protein property prediction, though protein structures are known to be determinants of protein function. In this paper, we propose to pretrain protein representations according to their 3D structures. We first present a simple yet effective encoder to learn the geometric features of a protein. We pretrain the protein graph encoder by leveraging multiview contrastive learning and different self-prediction tasks. Experimental results on both function prediction and fold classification tasks show that our proposed pretraining methods outperform or are on par with the state-of-the-art sequence-based methods, while using much less pretraining data. Our implementation is available at https://github.com/DeepGraphLearning/GearNet.
****

Conditional Invariances for Conformer Invariant Protein Representations
Balasubramaniam Srinivasan,Vassilis N. Ioannidis,Soji Adeshina,Mayank Kakodkar,George Karypis,Bruno Ribeiro

Representation learning for proteins is an emerging area in geometric deep learning. Recent works have factored in both the relational (atomic bonds) and the geometric aspects (atomic positions) of the task, notably bringing together graph neural networks (GNNs) with neural networks for point clouds. The equivariances and invariances to geometric transformations (group actions such as rotations and translations) so far treats large molecules as rigid structures. However, in many important settings, proteins can co-exist as an ensemble of multiple stable conformations. The conformations of a protein, however, cannot be described as input-independent transformations of the protein: Two proteins may require different sets of transformations in order to describe their set of viable conformations. To address this limitation, we introduce the concept of conditional transformations (CT). CT can capture protein structure, while respecting the restrictions posed by constraints on dihedral (torsion) angles and steric repulsions between atoms. We then introduce a Markov chain Monte Carlo framework to learn representations that are invariant to these conditional transformations. Our results show that endowing existing baseline models with these conditional transformations helps improve their performance without sacrificing computational cost.
****

Learning PDE Solution Operator for Continuous Modeling of Time-Series
Yesom Park,Jaemoo Choi,Changyeon Yoon,Chang hoon Song,Myungjoo Kang

Learning underlying dynamics from data is important and challenging in many real-world scenarios. Incorporating differential equations (DEs) to design continuous networks has drawn much attention recently, the most prominent of which is Neural ODE. Most prior works make specific assumptions on the type of DEs or restrict them to first or second-order DEs, making the model specialized for certain problems. Furthermore, due to the use of numerical integration, they suffer from computational expensiveness and numerical instability. Building upon recent Fourier neural operator (FNO), this work proposes a partial differential equation (PDE) based framework which improves the dynamics modeling capability and circumvents the need for costly numerical integration. FNO is hard to be directly applied to real applications because it is mainly confined to physical PDE problems. To fill this void, we propose a continuous-in-time FNO to deal with irregularly-sampled time series and provide a theoretical result demonstrating its universality. Moreover, we reveal an intrinsic property of PDEs that increases the stability of the model. Several numerical evidence shows that our method represents a b roader range of problems, including synthetic, image classification, and irregular time-series. Our framework opens up a new way for a continuous representation of neural networks that can be readily adopted for real-world applications.

****

Quantum-Inspired Tensorized Embedding with Application to Node Representation Learning

Hao Xiong,Yehui Tang,Wei Tan,Yunlin He,Junchi Yan

Node representation learning a.k.a. network embedding (NE) is an essential technique for network analysis by representing nodes as vectors, which also serves downstream tasks or as initial input for GNN models. Most of existing NE algorithms require a space complexity linear to the multiplication of the number of nodes and embedding dimension to store embeddings. Such a conventional embedding paradigm has two defects: i) it brings challenge to the deployment of NE algorithms for large-scale networks on devices of limited memory/storage space; ii) model expressiveness is constrained due to the limited embedding dimension. Impressed and inspired by the large Hilbert space of quantum systems, we propose a brand new NE algorithm \emph{node2ket} by imitating behaviors of quantum systems. Theoretically, we give analysis on how it unifies existing embedding methods including both conventional ones and tensorized ones, and explore the ultimate compressive power of the embedding model on the space complexity compared with conventional ones. Experiments are conducted on five public real-world networks where methods are evaluated through tasks of network reconstruction and link prediction. On BlogCatalog, our method achieves to outperform all baselines with 1/32 training parameters and 1/16 running time on the same machine. On DBLP, the reconstruction precision of node2ket achieves to be 3 times higher than the best baseline i.e. LouvainNE. Source code will be made publicly available.

****

Identifying Latent Causal Content for Multi-Source Domain Adaptation

Yuhang Liu,Zhen Zhang,Dong Gong,Mingming Gong,Biwei Huang,Kun Zhang,Javen Qinfeng Shi

Multi-source domain adaptation (MSDA) learns to predict the labels in target domain data, under the setting that data from multiple source domains are labelled and data from the target domain are unlabelled. Most methods for this task focus on learning invariant representations across domains. However, their success relies heavily on the assumption that the label distribution remains consistent across domains, which may not hold in general real-world problems. In this paper, we propose a new and more flexible assumption, termed \textit{latent covariate shift}, where a latent content variable $\mathbf{z}_c$ and a latent style variable $\mathbf{z}_s$ are introduced in the generative process, with the marginal distribution of $\mathbf{z}_c$ changing across domains and the conditional distribution of the label given $\mathbf{z}_c$ remaining invariant across domains. We show that although (completely) identifying the proposed latent causal model is challenging, the latent content variable can be identified up to scaling by using its dependence with labels from source domains, together with the identifiability conditions of nonlinear ICA. This motivates us to propose a novel method for MSDA, which learns the invariant label distribution conditional on the latent content variable, instead of learning invariant representations. Empirical evaluation on simulation and real data demonstrates the effectiveness of the proposed method.

****

Trainable Weight Averaging: Efficient Training by Optimizing Historical Solutions

Tao Li,Zhehao Huang,Qinghua Tao,Yingwen Wu,Xiaolin Huang

Stochastic gradient descent (SGD) and its variants are considered as the de-facto methods to train deep neural networks (DNNs). While recent improvements to SGD mainly focus on the descent algorithm itself, few works pay attention to utilizing the historical solutions---as an iterative method, SGD has gone through substantial explorations before convergence. Recently, an interesting attempt is stochastic weight averaging (SWA), which significantly improves the generalization by simply averaging the solutions at the tail stage of training. In this paper,

we realize that the averaging coefficients could be determined in a trainable manner and propose Trainable Weight Averaging (TWA), a novel optimization method in the reduced subspace spanned by historical solutions. TWA has much greater flexibility and can be applied to the head stage of training to achieve training efficiency while preserving good generalization capability. Further, we propose a distributed training scheme to resolve the memory burden of large-scale training with efficient parallel computation. In the extensive numerical experiments, (i) TWA achieves consistent improvements over SWA with less sensitivity to learning rate; (ii) applying TWA in the head stage of training largely speeds up the convergence, resulting in over 40% time saving on CIFAR and 30% on ImageNet with improved generalization compared with regular training.
****
Revealing Single Frame Bias for Video-and-Language Learning
Jie Lei,Tamara L Berg,Mohit Bansal
Training an effective video-and-language model intuitively requires multiple frames as model inputs.
However, it is unclear whether using multiple frames is beneficial to downstream tasks, and if yes, whether the performance gain is worth the drastically-increased computation and memory costs resulting from using more frames.
In this work, we explore single-frame models for video-and-language learning.
On a diverse set of video-and-language tasks (including text-to-video retrieval and video question answering), we show the surprising result that, with large-scale pre-training and a proper frame ensemble strategy at inference time, a single-frame trained model that does not consider temporal information can achieve better performance than existing methods that use multiple frames for training.
This result reveals the existence of a strong ``static appearance bias'' in popular video-and-language datasets.
Therefore, to allow for a more comprehensive evaluation of video-and-language models, we propose two new retrieval tasks based on existing fine-grained action recognition datasets that encourage temporal modeling.
Full code and models will be made publicly available upon acceptance.
****
Deep Declarative Dynamic Time Warping for End-to-End Learning of Alignment Paths
Ming Xu,Sourav Garg,Michael Milford,Stephen Gould
This paper addresses learning end-to-end models for time series data that include a temporal alignment step via dynamic time warping (DTW). Existing approaches to differentiable DTW either differentiate through a fixed warping path or apply a differentiable relaxation to the min operator found in the recursive steps used to solve the DTW problem. We instead propose a DTW layer based around bi-level optimisation and deep declarative networks, which we name DecDTW. By formulating DTW as a continuous, inequality constrained optimisation problem, we can compute gradients for the solution of the optimal alignment (with respect to the underlying time series) using implicit differentiation. An interesting byproduct of this formulation is that DecDTW outputs the optimal warping path between two time series as opposed to a soft approximation, recoverable from Soft-DTW. We show that this property is particularly useful for applications where downstream loss functions are defined on the optimal alignment path itself. This naturally occurs, for instance, when learning to improve the accuracy of predicted alignments against ground truth alignments. We evaluate DecDTW on two such applications, namely the audio-to-score alignment task in music information retrieval and the visual place recognition task in robotics, demonstrating state-of-the-art results in both.
****
DEEAPR: Controllable Depth Enhancement via Adaptive Parametric Feature Rotation
Hanul Shin,Youngchan Song,Soo Min Kang
Understanding depth of an image provides viewers with a better interpretation of the 3D structures within an image. Photographers utilize numerous factors that can affect depth perception to aesthetically improve a scene. Unfortunately, controlling depth perception after the image has been captured is a difficult process as it requires accurate and explicit depth information. Also, defining a quan

titative metric of a subjective quality (i.e., depth perception) is difficult which makes supervised learning a great challenge. To this end, we propose DEpth Enhancement via Adaptive Parametric feature Rotation (DEEAPR), which modulates the perceptual depth of an input scene using a single control parameter without the need for explicit depth information. We first embed content-independent depth perception of a scene by visual representation learning. Then, we train the controllable depth enhancer network with a novel modulator, parametric feature rotation block (PFRB), that allows for continuous modulation of a representative feature. We demonstrate the effectiveness of our proposed approach by verifying each component through an ablation study and comparison to other controllable methods.

****

Deep Active Anomaly Detection With Diverse Queries

Aodong Li,Chen Qiu,Padhraic Smyth,Marius Kloft,Stephan Mandt,Maja Rudolph

Selecting informative data points for expert feedback can significantly improve the performance of anomaly detection in various contexts, such as medical diagnostics or fraud detection. In this paper, we determine a set of conditions under which the ranking of anomaly scores generalizes from labeled queries to unlabeled data. Inspired by these conditions, we propose a new querying strategy for active anomaly detection that leads to systematic improvements over current approaches for this problem. It selects a diverse set of data points for labeling, achieving high data coverage with a limited budget. These labeled data points provide weak supervision to the unsupervised anomaly detection problem. However, correctly identifying anomalies requires an estimate of the fraction of anomalies in the data. We show how this anomaly rate can be estimated from the query set by importance-weighting, removing the associated bias due to the non-uniform sampling procedure. Extensive experiments on image, tabular, and video data sets show that our approach results in state-of-the-art active anomaly detection performance.

****

Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning

Ting Chen,Ruixiang ZHANG,Geoffrey Hinton

We present Bit Diffusion: a simple and generic approach for generating discrete data with continuous state and continuous time diffusion models. The main idea behind our approach is to first represent the discrete data as binary bits, and then train a continuous diffusion model to model these bits as real numbers which we call analog bits. To generate samples, the model first generates the analog bits, which are then thresholded to obtain the bits that represent the discrete variables. We further propose two simple techniques, namely Self-Conditioning and Asymmetric Time Intervals, which lead to a significant improvement in sample quality. Despite its simplicity, the proposed approach can achieve strong performance in both discrete image generation and image captioning tasks. For discrete image generation, we significantly improve previous state-of-the-art on both CIFAR-10 (which has 3K discrete 8-bit tokens) and ImageNet-64x64 (which has 12K discrete 8-bit tokens), outperforming the best autoregressive model in both sample quality (measured by FID) and efficiency. For image captioning on MS-COCO dataset, our approach achieves competitive results compared to autoregressive models.

****

Understanding Edge-of-Stability Training Dynamics with a Minimalist Example

Xingyu Zhu,Zixuan Wang,Xiang Wang,Mo Zhou,Rong Ge

Recently, researchers observed that gradient descent for deep neural networks operates in an ``edge-of-stability'' (EoS) regime: the sharpness (maximum eigenvalue of the Hessian) is often larger than stability threshold $2/\eta$ (where η is the step size). Despite this, the loss oscillates and converges in the long run, and the sharpness at the end is just slightly below $2/\eta$. While many other well-understood nonconvex objectives such as matrix factorization or two-layer networks can also converge despite large sharpness, there is often a larger gap between sharpness of the endpoint and $2/\eta$. In this paper, we study EoS phenomenon by constructing a simple function that has the same behavior. We giv e rigorous analysis for its training dynamics in a large local region and explain why the final converging point has sharpness close to $2/\eta$. Globally we observe that the training dynamics for our example has an interesting bifurcating behavior, which was also observed in the training of neural nets.

****

Learning Proximal Operators to Discover Multiple Optima

Lingxiao Li,Noam Aigerman,Vladimir Kim,Jiajin Li,Kristjan Greenewald,Mikhail Yurochkin,Justin Solomon

Finding multiple solutions of non-convex optimization problems is a ubiquitous yet challenging task. Most past algorithms either apply single-solution optimization methods from multiple random initial guesses or search in the vicinity of found solutions using ad hoc heuristics. We present an end-to-end method to learn the proximal operator of a family of training problems so that multiple local minima can be quickly obtained from initial guesses by iterating the learned operator, emulating the proximal-point algorithm that has fast convergence. The learned proximal operator can be further generalized to recover multiple optima for unseen problems at test time, enabling applications such as object detection. The key ingredient in our formulation is a proximal regularization term, which elevates the convexity of our training loss: by applying recent theoretical results, we show that for weakly-convex objectives with Lipschitz gradients, training of the proximal operator converges globally with a practical degree of over-parameterization. We further present an exhaustive benchmark for multi-solution optimization to demonstrate the effectiveness of our method.

****

Guiding continuous operator learning through Physics-based boundary constraints

Nadim Saad,Gaurav Gupta,Shima Alizadeh,Danielle C. Maddix

Boundary conditions (BCs) are important groups of physics-enforced constraints that are necessary for solutions of Partial Differential Equations (PDEs) to satisfy at specific spatial locations. These constraints carry important physical meaning, and guarantee the existence and the uniqueness of the PDE solution. Current neural-network based approaches that aim to solve PDEs rely only on training data to help the model learn BCs implicitly, however, there is no guarantee of BC satisfaction by these models during evaluation. In this work, we propose Boundary enforcing Operator Network (BOON) that enables the BC satisfaction of neural operators by making structural changes to the operator kernel. We provide our refinement procedure, and demonstrate the satisfaction of physics-based BCs such as Dirichlet, Neumann, and periodic by the solutions obtained by BOON. Numerical experiments based on multiple PDEs with a wide variety of applications indicate that the proposed approach ensures satisfaction of BCs, and leads to more accurate solutions over the whole domain. The proposed method exhibits a (2X-20X) improvement in accuracy (0.000084 relative L^2 error for Burgers' equation). Code available at: https://github.com/amazon-science/boon.

****

AdaWAC: Adaptively Weighted Augmentation Consistency Regularization for Volumetric Medical Image Segmentation

Yijun Dong,Yuege Xie,Rachel Ward

Sample reweighting is an effective strategy for learning from training data coming from a mixture of different subpopulations. However, existing reweighting algorithms do not fully take advantage of the particular type of data distribution encountered in volumetric medical image segmentation, where the training data images are uniformly distributed but their associated data labels fall into two subpopulations---"label-sparse" and "label-dense"---depending on whether the data image occurs near the beginning/end of the volumetric scan or the middle. For this setting, we propose AdaWAC as an adaptive weighting algorithm that assigns label-dense samples to supervised cross-entropy loss and label-sparse samples to unsupervised consistency regularization. We provide a convergence guarantee for AdaWAC by appealing to the theory of online mirror descent on saddle point problems. Moreover, we empirically demonstrate that AdaWAC not only enhances segmentation performance and sample efficiency but also improves robustness to the subpopulation shift in labels.

****

Limitations of the NTK for Understanding Generalization in Deep Learning

Nikhil Vyas,Yamini Bansal,Preetum Nakkiran

The “Neural Tangent Kernel” (NTK) (Jacot et al., 2018), and its empirical variants have been proposed as a proxy to capture certain behaviors of real neural networks. In this work, we study NTKs through the lens of scaling laws, and demonstrate that they fall short of explaining important aspects of neural network generalization. In particular, we demonstrate realistic settings where finite-width neural networks have significantly better data scaling exponents as compared to their corresponding empirical and infinite NTKs at initialization. This reveals a more fundamental difference between the real networks and NTKs, beyond just a few percentage points of test accuracy. Further, we show that even if the empirical NTK is allowed to be pre-trained on a constant number of samples, the kernel scaling does not catch up to the neural network scaling. Finally, we show that the empirical NTK continues to evolve throughout most of the training, in contrast with prior work which suggests that it stabilizes after a few epochs of training. Altogether, our work establishes concrete limitations of the NTK approach in understanding generalization of real networks on natural datasets.

****

Federated Learning of Large Models at the Edge via Principal Sub-Model Training

Yue Niu,Saurav Prakash,Souvik Kundu,Sunwoo Lee,Salman Avestimehr

Limited compute, memory, and communication capabilities of edge users create a significant bottleneck for federated learning (FL) of large models. Current literature typically tackles the challenge with a heterogeneous client setting or allows training to be offloaded to the server. However, the former requires a fraction of clients to train near-full models, which may not be achievable at the edge; while the latter can compromise privacy with sharing of intermediate representations or labels. In this work, we consider a realistic, but much less explored, cross-device FL setting in which no client has the capacity to train a full large model nor is willing to share any intermediate representations with the server. To this end, we present Principal Sub-Model (PriSM) training methodology, which leverages models’ low-rank structure and kernel orthogonality to train sub-models in the orthogonal kernel space. More specifically, by applying singular value decomposition to original kernels in the server model, PriSM first obtains a set of principal orthogonal kernels with importance weighed by their singular values. Thereafter, PriSM utilizes a novel sampling strategy that selects different subsets of the principal kernels independently to create sub-models for clients with reduced computation and communication requirements. Importantly, a kernel with a large singular value is assigned with a high sampling probability. Thus, each sub-model is a low-rank approximation of the full large model, and all clients together achieve nearly full coverage of the principal kernels. To further improve memory efficiency, PriSM exploits low-rank structure in intermediate representations and allows each sub-model to learn only a subset of them while still preserving training performance. Our extensive evaluations on multiple datasets in various resource-constrained settings demonstrate that PriSM can yield an improved performance of up to 10% compared to existing alternatives, when training sub-models with only 20% principal kernels ($\sim 5\%$ of the full server model.).

****

Implicit Offline Reinforcement Learning via Supervised Learning

Alexandre Piché,Rafael Pardinas,David Vazquez,Igor Mordatch,Igor Mordatch,Christopher Pal

Offline Reinforcement Learning (RL) via Supervised Learning is a simple and effective way to learn robotic skills from a dataset of varied behaviors. It is as simple as supervised learning and Behavior Cloning (BC) but takes advantage of the return information. On BC tasks, implicit models have been shown to match or outperform explicit ones. Despite the benefits of using implicit models to learn robotic skills via BC, Offline RL via Supervised Learning algorithms have been limited to explicit models. We show how implicit models leverage return information and match or outperform explicit algorithms to acquire robotic skills from fi xed datasets. Furthermore, we show how closely related our implicit methods are to other popular RL via Supervised Learning algorithms.

****

Augmentation Backdoors

Joseph Rance,Yiren Zhao,Ilia Shumailov,Robert D. Mullins

Data augmentation is used extensively to improve model generalisation. However, reliance on external libraries to implement augmentation methods introduces a vulnerability into the machine learning pipeline. It is well known that backdoors can be inserted into machine learning models through serving a modified dataset to train on. Augmentation therefore presents a perfect opportunity to perform this modification without requiring an initially backdoored dataset. In this paper we present three backdoor attacks that can be covertly inserted into data augmentation. Our attacks each insert a backdoor using a different type of computer vision augmentation transform, covering simple image transforms, GAN-based augmentation, and composition-based augmentation. By inserting the backdoor using these augmentation transforms, we make our backdoors difficult to detect, while still supporting arbitrary backdoor functionality. We evaluate our attacks on a range of computer vision benchmarks and demonstrate that an attacker is able to introduce backdoors through just a malicious augmentation routine.

****

Neural Radiance Field Codebooks

Matthew Wallingford,Aditya Kusupati,Alex Fang,Vivek Ramanujan,Aniruddha Kembhavi,Roozbeh Mottaghi,Ali Farhadi

Compositional representations of the world are a promising step towards enabling high-level scene understanding and efficient transfer to downstream tasks. Learning such representations for complex scenes and tasks remains an open challenge. Towards this goal, we introduce Neural Radiance Field Codebooks (NRC), a scalable method for learning object-centric representations through novel view reconstruction. NRC learns to reconstruct scenes from novel views using a dictionary of object codes which are decoded through a volumetric renderer. This enables the discovery of reoccurring visual and geometric patterns across scenes which are transferable to downstream tasks. We show that NRC representations transfer well to object navigation in THOR, outperforming 2D and 3D representation learning methods by 3.1\% success rate. We demonstrate that our approach is able to perform unsupervised segmentation for more complex synthetic (THOR) and real scenes (NYU Depth) better than prior methods (.101 ARI). Finally, we show that NRC improves on the task of depth ordering by 5.5% accuracy in THOR.

****

Generalized Precision Matrix for Scalable Estimation of Nonparametric Markov Networks

Yujia Zheng,Ignavier Ng,Yewen Fan,Kun Zhang

A Markov network characterizes the conditional independence structure, or Markov property, among a set of random variables. Existing work focuses on specific families of distributions (e.g., exponential families) and/or certain structures of graphs, and most of them can only handle variables of a single data type (continuous or discrete). In this work, we characterize the conditional independence structure in general distributions for all data types (i.e., continuous, discrete, and mixed-type) with a Generalized Precision Matrix (GPM). Besides, we also allow general functional relations among variables, thus giving rise to a Markov network structure learning algorithm in one of the most general settings. To deal with the computational challenge of the problem, especially for large graphs, we unify all cases under the same umbrella of a regularized score matching framework. We validate the theoretical results and demonstrate the scalability empirically in various settings.

****

Determinant regularization for Deep Metric Learning

Kun Song,Ruben Solozabal,Martin Taká■,Fakhri Karray

Distance Metric Learning (DML) aims to learn the distance metric that better reflects the semantical similarities in the data. Current \textit{pair-based} and \textit{proxy-based} methods on DML focus on reducing the distance between simila r samples while expanding the distance of dissimilar ones. However, we reveal that shrinking the distance between similar samples may distort the feature space, increasing the distance between points of the same class region and, therefore, harming the generalization of the model. Traditional regularization terms (such as L_2-norm on weights) cannot be adopted to solve this issue as they are based on linear projection. To alleviate this issue, we adopt the structure of normalizing flow as the deep metric layer and calculate the determinant of the Jacobi Matrix as the regularization term. At last, we conduct experiments on several \textit{pair-based} and \textit{proxy-based} algorithms that demonstrate the benefits of our method.

****

Data-Efficient and Interpretable Tabular Anomaly Detection

Chun-Hao Chang,Jinsung Yoon,Sercan O Arik,Madeleine Udell,Tomas Pfister

Anomaly detection (AD) plays an important role in numerous applications. In this paper, we focus on two understudied aspects of AD that are critical for integration into real-world applications. First, most AD methods cannot incorporate labeled data that are often available in practice in small quantities and can be crucial to achieve high accuracy. Second, most AD methods are not interpretable, a bottleneck that prevents stakeholders from understanding the reason behind the anomalies. In this paper, we propose a novel AD framework, DIAD, that adapts a white-box model class, Generalized Additive Models, to detect anomalies using a partial identification objective which naturally handles noisy or heterogeneous features. DIAD can incorporate a small amount of labeled data to further boost AD performances in semi-supervised settings. We demonstrate the superiority of DIAD compared to previous work in both unsupervised and semi-supervised settings on multiple datasets. We also present explainability capabilities of DIAD, on its rationale behind predicting certain samples as anomalies.

****

FiT: Parameter Efficient Few-shot Transfer Learning for Personalized and Federated Image Classification

Aliaksandra Shysheya,John F Bronskill,Massimiliano Patacchiola,Sebastian Nowozin,Richard E Turner

Modern deep learning systems are increasingly deployed in situations such as personalization and federated learning where it is necessary to support i) learning on small amounts of data, and ii) communication efficient distributed training protocols. In this work, we develop FiLM Transfer (FiT) which fulfills these requirements in the image classification setting by combining ideas from transfer learning (fixed pretrained backbones and fine-tuned FiLM adapter layers) and meta-learning (automatically configured Naive Bayes classifiers and episodic training) to yield parameter efficient models with superior classification accuracy at low-shot. The resulting parameter efficiency is key for enabling few-shot learning, inexpensive model updates for personalization, and communication efficient federated learning. We experiment with FiT on a wide range of downstream datasets and show that it achieves better classification accuracy than the leading Big Transfer (BiT) algorithm at low-shot and achieves state-of-the art accuracy on the challenging VTAB-1k benchmark, with fewer than 1% of the updateable parameters. Finally, we demonstrate the parameter efficiency and superior accuracy of FiT in distributed low-shot applications including model personalization and federated learning where model update size is an important performance metric.

****

A Critical Analysis of Out-of-Distribution Detection for Document Understanding

Jiuxiang Gu,Yifei Ming,Yi Zhou,Jason Kuen,Vlad I Morariu,Handong Zhao,Ruiyi Zhang,Nikolaos Barmpalios,Anqi Liu,Yixuan Li,Tong Sun,Ani Nenkova

Large-scale pretraining is widely used in recent document understanding models. During deployment, one may expect that large-scale pretrained models should trigger a conservative fallback policy when encountering out-of-distribution (OOD) samples, which suggests the importance of OOD detection. However, most existing OOD detection methods focus on single-modal inputs such as images or texts. While

documents are multi-modal in nature, it is underexplored if and how multi-modal information in documents can be exploited for OOD detection. In this work, we first provide a systematic and in-depth analysis on OOD detection for document understanding models. We study the effects of model modality, pretraining, and finetuning across various types of OOD inputs. In particular, we find that spatial information is critical for document OOD detection. To better exploit spatial information, we propose a simple yet effective special-aware adapter, which serves as an add-on module to adapt transformer-based language models to document domain. Extensive experiments show that our method consistently improves ID accuracy and OOD detection performance compared to baselines. We hope our findings can help inspire future works on understanding OOD robustness for documents.
****

Learnable Visual Words for Interpreting Image Recognition Models
Wenxiao Xiao,Zhengming Ding,Hongfu Liu

To interpret deep models' predictions, attention-based visual cues are widely used in addressing *why* deep models make such predictions. Beyond that, the current research community becomes more interested in reasoning *how* deep models make predictions, where some prototype-based methods employ interpretable representations with their corresponding visual cues to reveal the black-box mechanism of deep model behaviors. However, these pioneering attempts only either learn the category-specific prototypes and deteriorate with their generalization ability deterioration or demonstrate several illustrative examples without a quantitative evaluation of visual-based interpretability narrowing their practical usages. In this paper, we revisit the concept of visual words and propose the Learnable Visual Words (LVW) to interpret the model prediction behaviors with two novel modules: semantic visual words learning and dual fidelity preservation. The semantic visual words learning relaxes the category-specific constraint, enabling the generic visual words shared across multiple categories. Beyond employing the visual words for prediction to align visual words with the base model, our dual fidelity preservation also includes the attention-guided semantic alignment that encourages the learned visual words to focus on the same conceptual regions for prediction. Experiments on six visual benchmarks demonstrate the superior effectiveness of our proposed LVW in both accuracy and interpretation fidelity over the state-of-the-art methods. Moreover, we elaborate on various in-depth analyses to further explore the learned visual words and the generalizability of our method for unseen categories.
****

Compact Bilinear Pooling via General Bilinear Projection
Kun Song,Junwei Han,Feiping Nie,Gong Cheng,Bin Gu,Fakhri Karray

Most factorized bilinear pooling (FBiP) employs Hadamard product-based bilinear projection to learn appropriate projecting directions to reduce the dimension of bilinear features. However, in this paper, we reveal that the Hadamard product-based bilinear projection makes FBiP miss a lot of possible projecting directions, which will significantly harm the performance of outputted compact bilinear features, including compactness and effectiveness. To address this issue, we propose a general matrix-based bilinear projection based on the rank-k matrix base decomposition, where the Hadamard-based bilinear projection is a special case of our proposed one. Using the proposed bilinear projection, we design a novel low-rank factorized bilinear pooling (named RK-FBP), which does not miss any projecting directions. Thus, our RK-FBP can generate better compact bilinear features. To leverage high-order information in local features, we nest several RK-FBP modules together to formulate a multi-linear pooling that outputs compact multi-linear features. At last, we conduct experiments on several fine-grained image tasks to evaluate our models. The experiments show that our models achieve new state-of-the-art classification accuracy by the lowest dimension.
****

AANG : Automating Auxiliary Learning
Lucio M. Dery,Paul Michel,Mikhail Khodak,Graham Neubig,Ameet Talwalkar

Auxiliary objectives, supplementary learning signals that are introduced to help aid learning on data-starved or highly complex end-tasks, are commonplace in ma chine learning. Whilst much work has been done to formulate useful auxiliary objectives, their construction is still an art which proceeds by slow and tedious hand-design. Intuition for how and when these objectives improve end-task performance has also had limited theoretical backing. In this work, we present an approach for automatically generating a suite of auxiliary objectives. We achieve this by deconstructing existing objectives within a novel unified taxonomy, identifying connections between them, and generating new ones based on the uncovered structure. Next, we theoretically formalize widely-held intuitions about how auxiliary learning improves generalization on the end-task. This leads us to a principled and efficient algorithm for searching the space of generated objectives to find those most useful to a specified end-task.

With natural language processing (NLP) as our domain of study, we demonstrate that our automated auxiliary learning pipeline leads to strong improvements over competitive baselines across continued training experiments on a pre-trained model on 5 NLP end-tasks.

****

Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation

Ye Zhu,Yu Wu,Kyle Olszewski,Jian Ren,Sergey Tulyakov,Yan Yan

Diffusion probabilistic models (DPMs) have become a popular approach to conditional generation, due to their promising results and support for cross-modal synthesis. A key desideratum in conditional synthesis is to achieve high correspondence between the conditioning input and generated output. Most existing methods learn such relationships implicitly, by incorporating the prior into the variational lower bound. In this work, we take a different route---we explicitly enhance input-output connections by maximizing their mutual information. To this end, we introduce a Conditional Discrete Contrastive Diffusion (CDCD) loss and design two contrastive diffusion mechanisms to effectively incorporate it into the denoising process, combining the diffusion training and contrastive learning for the first time by connecting it with the conventional variational objectives. We demonstrate the efficacy of our approach in evaluations with diverse multimodal conditional synthesis tasks: dance-to-music generation, text-to-image synthesis, as well as class-conditioned image synthesis. On each, we enhance the input-output correspondence and achieve higher or competitive general synthesis quality. Furthermore, the proposed approach improves the convergence of diffusion models, reducing the number of required diffusion steps by more than 35% on two benchmarks, significantly increasing the inference speed.

****

Diffusion Probabilistic Modeling of Protein Backbones in 3D for the motif-scaffolding problem

Brian L. Trippe,Jason Yim,Doug Tischer,David Baker,Tamara Broderick,Regina Barzilay,Tommi S. Jaakkola

Construction of a scaffold structure that supports a desired motif, conferring protein function, shows promise for the design of vaccines and enzymes. But a general solution to this motif-scaffolding problem remains open. Current machine-learning techniques for scaffold design are either limited to unrealistically small scaffolds (up to length 20) or struggle to produce multiple diverse scaffolds. We propose to learn a distribution over diverse and longer protein backbone structures via an E(3)-equivariant graph neural network. We develop SMCDiff to efficiently sample scaffolds from this distribution conditioned on a given motif; our algorithm is the first to theoretically guarantee conditional samples from a diffusion model in the large-compute limit. We evaluate our designed backbones by how well they align with AlphaFold2-predicted structures. We show that our method can (1) sample scaffolds up to 80 residues and (2) achieve structurally diverse scaffolds for a fixed motif.

****

NeRF-SOS: Any-View Self-supervised Object Segmentation on Complex Scenes

Zhiwen Fan,Peihao Wang,Yifan Jiang,Xinyu Gong,Dejia Xu,Zhangyang Wang

Neural volumetric representations have shown the potential that Multi-layer Perceptrons (MLPs) can be optimized with multi-view calibrated images to represent scene geometry and appearance without explicit 3D supervision. Object segmentatio n can enrich many downstream applications based on the learned radiance field. However, introducing hand-crafted segmentation to define regions of interest in a complex real-world scene is non-trivial and expensive as it acquires per view annotation. This paper carries out the exploration of self-supervised learning for object segmentation using NeRF for complex real-world scenes. Our framework, called NeRF with Self-supervised Object Segmentation (NeRF-SOS), couples object segmentation and neural radiance field to segment objects in any view within a scene. By proposing a novel collaborative contrastive loss in both appearance and geometry levels, NeRF-SOS encourages NeRF models to distill compact geometry-aware segmentation clusters from their density fields and the self-supervised pre-trained 2D visual features. The self-supervised object segmentation framework can be applied to various NeRF models that both lead to photo-realistic rendering results and convincing segmentation maps for both indoor and outdoor scenarios. Extensive results on the LLFF, BlendedMVS, CO3Dv2, and Tank & Temples datasets validate the effectiveness of NeRF-SOS. It consistently surpasses other 2D-based self-supervised baselines and predicts finer object masks than existing supervised counterparts.

****

RbX: Region-based explanations of prediction models
Ismael Lemhadri,Harrison H Li,Trevor Hastie

We introduce region-based explanations (RbX), a novel, model-agnostic method to generate local explanations of scalar outputs from a black-box prediction model using only query access. RbX is based on a greedy algorithm for building a convex polytope that approximates a region of feature space where model predictions are close to the prediction at some target point. This region is fully specified by the user on the scale of the predictions, rather than on the scale of the features. The geometry of this polytope - specifically the change in each coordinate necessary to escape the polytope - quantifies the local sensitivity of the predictions to each of the features. These “escape distances” can then be standardized to rank the features by local importance. RbX is guaranteed to satisfy a “sparsity” axiom, which requires that features which do not enter into the prediction model are assigned zero importance. At the same time, real data examples and synthetic experiments show how RbX can more readily detect all locally relevant features than existing methods.

****

Rethinking Graph Lottery Tickets: Graph Sparsity Matters
Bo Hui,Da Yan,Xiaolong Ma,Wei-Shinn Ku

Lottery Ticket Hypothesis (LTH) claims the existence of a winning ticket (i.e., a properly pruned sub-network together with original weight initialization) that can achieve competitive performance to the original dense network. A recent work, called UGS, extended LTH to prune graph neural networks (GNNs) for effectively accelerating GNN inference. UGS simultaneously prunes the graph adjacency matrix and the model weights using the same masking mechanism, but since the roles of the graph adjacency matrix and the weight matrices are very different, we find that their sparsifications lead to different performance characteristics. Specifically, we find that the performance of a sparsified GNN degrades significantly when the graph sparsity goes beyond a certain extent. Therefore, we propose two techniques to improve GNN performance when the graph sparsity is high. First, UGS prunes the adjacency matrix using a loss formulation which, however, does not properly involve all elements of the adjacency matrix; in contrast, we add a new auxiliary loss head to better guide the edge pruning by involving the entire adjacency matrix. Second, by regarding unfavorable graph sparsification as adversarial data perturbations, we formulate the pruning process as a min-max optimization problem to gain the robustness of lottery tickets when the graph sparsity is high. We further investigate the question: Can the ``retrainable'' winning ticket of a GNN be also effective for graph transferring learning? We call it the transferable graph lottery ticket (GLT) hypothesis. Extensive experiments were conducted which demonstrate the superiority of our proposed sparsification method over UGS, and which empirically verified our transferable GLT hypothesis.

****

The Impact of Approximation Errors on Warm-Start Reinforcement Learning: A Finite-time Analysis
Hang Wang,Sen Lin,Junshan Zhang

Warm-Start reinforcement learning (RL), aided by a prior policy obtained from offline training, is emerging as a promising RL approach for practical applications. Recent empirical studies have demonstrated that the performance of Warm-Start RL can be improved \textit{quickly} in some cases but become \textit{stagnant} in other cases, calling for a fundamental understanding, especially when the function approximation is used. To fill this void, we take a finite time analysis approach to quantify the impact of approximation errors on the learning performance of Warm-Start RL. Specifically, we consider the widely used Actor-Critic (A-C) method with a prior policy. We first quantify the approximation errors in the Actor update and the Critic update, respectively. Next, we cast the Warm-Start A-C algorithm   as Newton's method with perturbation, and study the impact of the approximation errors on the finite-time learning performance with inaccurate Actor/Critic updates. Under some general technical conditions, we obtain lower bounds on the sub-optimality gap of the Warm-Start A-C algorithm to quantify the impact of the bias and error propagation. We also derive the upper bounds, which provide insights on achieving the desired finite-learning performance in the Warm-Start A-C algorithm.

****

NeRN: Learning Neural Representations for Neural Networks
Maor Ashkenazi,Zohar Rimon,Ron Vainshtein,Shir Levi,Elad Richardson,Pinchas Mintz,Eran Treister

Neural Representations have recently been shown to effectively reconstruct a wide range of signals from 3D meshes and shapes to images and videos. We show that, when adapted correctly, neural representations can be used to directly represent the weights of a pre-trained convolutional neural network, resulting in a Neural Representation for Neural Networks (NeRN). Inspired by coordinate inputs of previous neural representation methods, we assign a coordinate to each convolutional kernel in our network based on its position in the architecture, and optimize a predictor network to map coordinates to their corresponding weights. Similarly to the spatial smoothness of visual scenes, we show that incorporating a smoothness constraint over the original network's weights aids NeRN towards a better reconstruction. In addition, since slight perturbations in pre-trained model weights can result in a considerable accuracy loss, we employ techniques from the field of knowledge distillation to stabilize the learning process. We demonstrate the effectiveness of NeRN in reconstructing widely used architectures on CIFAR-10, CIFAR-100, and ImageNet. Finally, we present two applications using NeRN, demonstrating the capabilities of the learned representations.

****

Private Federated Learning Without a Trusted Server: Optimal Algorithms for Convex Losses
Andrew Lowy,Meisam Razaviyayn

This paper studies federated learning (FL)—especially cross-silo FL—with data from people who do not trust the server or other silos. In this setting, each silo (e.g. hospital) has data from different people (e.g. patients) and must maintain the privacy of each person’s data (e.g. medical record), even if the server or other silos act as adversarial eavesdroppers. This requirement motivates the study of Inter-Silo Record-Level Differential Privacy (ISRL-DP), which requires silo i’s communications to satisfy record/item-level differential privacy (DP). ISRL-DP ensures that the data of each person (e.g. patient) in silo i (e.g. hospital i) cannot be leaked. ISRL-DP is different from well-studied privacy notions. Central and user-level DP assume that people trust the server/other silos. On the other end of the spectrum, local DP assumes that people do not trust anyone at all (even their own silo). Sitting between central and local DP, ISRL-DP makes the realistic assumption (in cross-silo FL) that people trust their own silo, but not the server or other silos. In this work, we provide tight (up to logarithms) upper and lower bounds for ISRL-DP FL with convex/strongly convex loss functions and homogeneous (i.i.d.) silo data. Remarkably, we show that similar b

ounds are attainable for smooth losses with arbitrary heterogeneous silo data distributions, via an accelerated ISRL-DP algorithm. We also provide tight upper and lower bounds for ISRL-DP federated empirical risk minimization, and use acceleration to attain the optimal bounds in fewer rounds of communication than the state-of-the-art. Finally, with a secure “shuffler” to anonymize silo messages (but without a trusted server), our algorithm attains the optimal central DP rates under more practical trust assumptions. Numerical experiments show favorable privacy-accuracy tradeoffs for our algorithm in classification and regression tasks.
****

3D-Aware Video Generation
Sherwin Bahmani,Jeong Joon Park,Despoina Paschalidou,Hao Tang,Gordon Wetzstein,Leonidas Guibas,Luc Van Gool,Radu Timofte
Generative models have emerged as an essential building block for many image synthesis and editing tasks. Recent advances in this field have also enabled high-quality 3D or video content to be generated that exhibits either multi-view or temporal consistency. With our work, we explore 4D generative adversarial networks (GANs) that learn unconditional generation of 3D-aware videos. By combining neural implicit representations with time-aware discriminator, we develop a GAN framework that synthesizes 3D video supervised only with monocular videos. We show that our method learns a rich embedding of decomposable 3D structures and motions that enables new visual effects of spatio-temporal renderings while producing imagery with quality comparable to that of existing 3D or video GANs.
****

Joint rotational invariance and adversarial training of a dual-stream Transformer yields state of the art Brain-Score for Area V4
William Berrios,Arturo Deza
Modern high-scoring models of vision in the brain score competition do not stem from Vision Transformers. However, in this paper, we provide evidence against the unexpected trend of Vision Transformers (ViT) being not perceptually aligned with human visual representations by showing how a dual-stream Transformer, a CrossViT $~\textit{a la}$ Chen et. al. (2021), under a joint rotationally-invariant and adversarial optimization procedure yields 2nd place in the aggregate Brain-Score 2022 competition (Schrimpf et al., 2020b) averaged across all visual categories, and at the time of the competition held 1st place for the highest explainable variance of area V4. In addition, our current Transformer-based model also achieves greater explainable variance for areas V4, IT, and Behaviour than a biologically-inspired CNN (ResNet50) that integrates a frontal V1-like computation module (Dapello et al., 2020). To assess the contribution of the optimization scheme with respect to the CrossViT architecture, we perform several additional experiments on differently optimized CrossViT's regarding adversarial robustness, common corruption benchmarks, mid-ventral stimuli interpretation, and feature inversion. Against our initial expectations, our family of results provides tentative support for an $\textit{``All roads lead to Rome''}$ argument enforced via a joint optimization rule even for non biologically-motivated models of vision such as Vision Transformers.
****

AutoSparse: Towards Automated Sparse Training
Abhisek Kundu,Naveen Mellempudi,Dharma Teja Vooturi,Bharat Kaul,Pradeep Dubey
Sparse training is emerging as a promising avenue for reducing the computational cost of training neural networks. Several recent studies have proposed pruning methods using learnable thresholds to efficiently explore the non-uniform distribution of sparsity inherent within the models. In this paper, we propose Gradient Annealing (GA), a gradient driven approach where gradients to pruned out weights are scaled down in a non-linear manner. GA eliminates the need for additional sparsity-inducing regularization by providing an elegant trade-off between sparsity and accuracy. We integrated GA with the latest learnable threshold based pruning methods to create an automated sparse training algorithm called AutoSparse. Our algorithm achieves state-of-the-art accuracy with 80% sparsity for ResNet50 and 75% sparsity for MobileNetV1 on Imagenet-1K. AutoSparse also results in 7

× reduction in inference FLOPS and > 2× reduction in training FLOPS for ResNet50 on ImageNet at 80% sparsity. Finally, GA generalizes well to fixed-budget (Top-K, 80%) sparse training methods, improving the accuracy of ResNet50 on Imagenet-1K, to outperform TopKAST+PP by 0.3%.
****

Learning to Communicate using Contrastive Learning
Yat Long Lo,Biswa Sengupta,Jakob Nicolaus Foerster,Michael Noukhovitch
Communication is a powerful tool for coordination in multi-agent RL. Inducing an effective, common language has been 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. Based on this perspective, we propose to learn to communicate using contrastive learning by maximizing 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 task-relevant information from the environment. Finally, we demonstrate promising results on zero-shot communication, a first for MARL. Overall, we show the power of contrastive learning, and self-supervised learning in general, as a method for learning to communicate.
****

Cheap Talk Discovery and Utilization in Multi-Agent Reinforcement Learning
Yat Long Lo,Christian Schroeder de Witt,Samuel Sokota,Jakob Nicolaus Foerster,Shimon Whiteson
By enabling agents to communicate, recent cooperative multi-agent reinforcement learning (MARL) methods have demonstrated better task performance and more coordinated behavior. Most existing approaches facilitate inter-agent communication by allowing agents to send messages to each other through free communication channels, i.e., \emph{cheap talk channels}. Current methods require these channels to be constantly accessible and known to the agents a priori. In this work, we lift these requirements such that the agents must discover the cheap talk channels and learn how to use them. Hence, the problem has two main parts: \emph{cheap talk discovery} (CTD) and \emph{cheap talk utilization} (CTU). We introduce a novel conceptual framework for both parts and develop a new algorithm based on mutual information maximization that outperforms existing algorithms in CTD/CTU settings. We also release a novel benchmark suite to stimulate future research in CTD/CTU.
****

Motif-induced Graph Normalization
Kaixuan Chen,Shunyu Liu,Tongtian Zhu,Tongya Zheng,Haofei Zhang,Jie Song,Mingli Song
Graph Neural Networks (GNNs) have emerged as a powerful category of learning architecture for handling graph-structured data in the non-Euclidean domain. Despite their success, existing GNNs typically suffer from the insufficient expressive power bottlenecked by Weisfeiler-Lehman (WL) test, and meanwhile are prone to the over-smoothing situation with increasing layer numbers. In this paper, we strive to strengthen the discriminative capabilities of GNNs by devising a dedicated plug-and-play normalization scheme, termed as Motif-induced Normalization (MotifNorm), that explicitly considers the intra-connection information within each node-induced subgraph. To this end, we embed the motif-induced structural weights at the beginning and the end of the standard BatchNorm, as well as incorporate the graph instance-specific statistics for improved distinguishable capabilities. In the meantime, we provide the theoretical analysis to support that, with the proposed elaborated MotifNorm, an arbitrary GNNs is capable of more expressive abilities than the 1-WL test in distinguishing k-regular graphs. Furthermore, the proposed MotifNorm scheme is also exemplified to be able to alleviate the over-smoothing phenomenon. Experimental results on ten popular benchmarks across all the tasks of the graph-, node-, as well as link-level property predictions, demonstrate the effectiveness of the proposed method. Our code is made available in the supplementary material.

****

Stochastic Gradient Methods with Preconditioned Updates

Abdurakhmon Sadiev,Aleksandr Beznosikov,Dmitry Kamzolov,Rachael Tappenden,Martin Taká■

This work considers non-convex finite sum minimization. There are a number of algorithms for such problems, but existing methods often work poorly when the problem is badly scaled and/or ill-conditioned, and a primary goal of this work is to introduce methods that alleviate this issue. Thus, here we include a preconditioner that is based upon Hutchinson's approach to approximating the diagonal of the Hessian, and couple it with several gradient based methods to give new `scaled' algorithms: Scaled SARAH and Scaled L-SVRG. Theoretical complexity guarantees under smoothness assumptions are presented, and we prove linear convergence when both smoothness and the PL-condition is assumed. Because our adaptively scaled methods use approximate partial second order curvature information, they are better able to mitigate the impact of badly scaled problems, and this improved practical performance is demonstrated in the numerical experiments that are also presented in this work.

****

Reversible Column Networks

Yuxuan Cai,Yizhuang Zhou,Qi Han,Jianjian Sun,Xiangwen Kong,Jun Li,Xiangyu Zhang

We propose a new neural network design paradigm Reversible Column Network (RevCol). The main body of RevCol is composed of multiple copies of subnetworks, named columns respectively, between which multi-level reversible connections are employed. Such architectural scheme attributes RevCol very different behavior from conventional networks: during forward propagation, features in RevCol are learned to be gradually disentangled when passing through each column, whose total information is maintained rather than compressed or discarded as other network does. Our experiments suggest that CNN-style RevCol models can achieve very competitive performances on multiple computer vision tasks such as image classification, object detection and semantic segmentation, especially with large parameter budget and large dataset. For example, after ImageNet-22K pre-training, RevCol-XL obtains 88.2% ImageNet-1K accuracy. Given more pre-training data, our largest model RevCol-H reaches 90.0% on ImageNet-1K, 63.8% AP$_{box}$ on COCO detection minival set, 61.0% mIoU on ADE20k segmentation. To our knowledge, it is the best COCO detection and ADE20k segmentation result among pure (static) CNN models. Moreover, as a general macro architecture fashion, RevCol can also be introduced into transformers or other neural networks, which is demonstrated to improve the performances in both computer vision and NLP tasks.
We release code and models at https://github.com/megvii-research/RevCol

****

Formal Mathematics Statement Curriculum Learning

Stanislas Polu,Jesse Michael Han,Kunhao Zheng,Mantas Baksys,Igor Babuschkin,Ilya Sutskever

We explore the use of expert iteration in the context of language modeling applied to formal mathematics. We show that at same compute budget, expert iteration, by which we mean proof search interleaved with learning, dramatically outperforms proof search only. We also observe that when applied to a collection of formal statements of sufficiently varied difficulty, expert iteration is capable of finding and solving a curriculum of increasingly difficult problems, without the need for associated ground-truth proofs. Finally, by applying this expert iteration to a manually curated set of problem statements, we surpass previous state-of-the-art on the miniF2F benchmark, automatically solving multiple challenging problems drawn from high school olympiads.

****

A Unified Causal View of Domain Invariant Representation Learning

Zihao Wang,Victor Veitch

Machine learning methods can be unreliable when deployed in domains that differ from the domains on which they were trained. One intuitive approach for addressing this is to learn representations of data that are domain-invariant in the sense that they preserve data structure that is stable across domains, but throw ou t spuriously-varying parts. There are many approaches aimed at this kind of representation-learning, including methods based on data augmentation, distributional invariances, and risk invariance. Unfortunately, it is often unclear when a given method actually learns domain-invariant structure, and whether learning domain-invariant structure actually yields robust models. The key issue is that, in general, it's unclear how to formalize ``domain-invariant structure''. The purpose of this paper is to study these questions in the context of a particular natural domain shift notion that admits a natural formal notion of domain invariance.

This notion is a formalization of the idea that causal relationships are invariant, but non-causal relationships (e.g., due to confounding) may vary. We find that whether a given method learns domain-invariant structure, and whether this leads to robust prediction, both depend critically on the true underlying causal structure of the data.

****

PIPS: Path Integral Stochastic Optimal Control for Path Sampling in Molecular Dynamics

Lars Holdijk,Yuanqi Du,Ferry Hooft,Priyank Jaini,Bernd Ensing,Max Welling

We consider the problem of Sampling Transition Paths: Given two metastable conformational states of a molecular system, \eg\ a folded and unfolded protein, we aim to sample the most likely transition path between the two states. Sampling such a transition path is computationally expensive due to the existence of high free energy barriers between the two states. To circumvent this, previous work has focused on simplifying the trajectories to occur along specific molecular descriptors called Collective Variables (CVs). However, finding CVs is non trivial and requires chemical intuition. For larger molecules, where intuition is not sufficient, using these CV-based methods biases the transition along possibly irrelevant dimensions. In this work, we propose a method for sampling transition paths that considers the entire geometry of the molecules. We achieve this by relating the problem to recent works on the Schr\"odinger bridge problem and stochastic optimal control. Using this relation, we construct a path integral method that incorporates important characteristics of molecular systems such as second-order dynamics and invariance to rotations and translations. We demonstrate our method on commonly studied protein structures like Alanine Dipeptide, and also consider larger proteins such as Polyproline and Chignolin.

****

ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets

Damien Teney,LIN Yong,Seong Joon Oh,Ehsan Abbasnejad

Several studies have empirically compared in-distribution (ID) and out-of-distribution (OOD) performance of various models. They report frequent positive correlations on benchmarks in computer vision and NLP. Surprisingly, they never observe inverse correlations suggesting necessary trade-offs. This matters to determine whether ID performance can serve as a proxy for OOD generalization. This paper shows that inverse correlations between ID and OOD performance do happen in real-world benchmarks. They could be missed in past studies because of a biased selection of models. We show an example on the WILDS-Camelyon17 dataset, using models from multiple training epochs and random seeds. Our observations are particularly striking with models trained with a regularizer that diversifies the solutions to the ERM objective. We nuance recommendations and conclusions made in past studies. (1) High OOD performance may sometimes require trading off ID performance.(2) Focusing on ID performance alone may not lead to optimal OOD performance: it can lead to diminishing and eventually negative returns in OOD performance. (3) Our example reminds that empirical studies only chart regimes achievable with existing methods: care is warranted in deriving prescriptive recommendations.

****

Visual Transformation Telling

Xin Hong,Yanyan Lan,Liang Pang,Jiafeng Guo,Xueqi Cheng

In this paper, we propose a new visual reasoning task, called Visual Transformation Telling (VTT). Given a series of states (i.e.~images), a machine is required to describe what happened (i.e.~transformation) between every two adjacent stat es. Different from most existing visual reasoning tasks, which focus on state reasoning, VTT concentrates on transformation reasoning. Moreover, describing the transformation in the form of language is more natural and closer to the real application than the property change way in the previous TVR task. We collect 13,547 samples from two instructional video datasets, i.e.~CrossTask and COIN, and extract desired states and transformation descriptions to form a suitable VTT benchmark dataset. After that, we introduce an end-to-end learning model for VTT, named TTNet. TTNet consists of three components to mimic human's cognition process of reasoning transformation. First, an image encoder, e.g. CLIP, reads content from each image, then a context encoder links the image content together, and at last, a transformation decoder autoregressively generates transformation descriptions between every two adjacent images. This basic version of TTNet is difficult to meet the cognitive challenge of VTT, that is to identify abstract transformations from images with small visual differences, and the descriptive challenge, which asks to describe the transformation consistently. In response to these difficulties, we propose three strategies to improve TTNet. Specifically, TTNet leverages difference features to emphasize small visual gaps, masked transformation model to stress context by forcing attention to neighbor transformations, and auxiliary category and topic classification tasks to make transformations consistent by sharing underlying semantics among representations. We adapt some typical methods from visual storytelling and dense video captioning tasks, considering their similarity with VTT. Our experimental results show that TTNet achieves better performance on transformation reasoning. In addition, our empirical analysis demonstrates the soundness of each module in TTNet, and provides some insight into transformation reasoning.

****

Predicting Antimicrobial MICs for Nontyphoidal Salmonella Using Multitask Representations Learning

Teng Lin

The antimicrobial resistance (AMR) pathogens have become an increasingly worldwide issue, posing a significant threat to global public health. To obtain an optimized therapeutic effect, the antibiotic sensitivity is usually evaluated in a clinical setting, whereas traditional culture-dependent antimicrobial sensitivity tests are labor-intensive and relatively slow. Rapid methods can greatly optimize antimicrobial therapeutic strategies and improve patient outcomes by reducing the time it takes to test antibiotic sensitivity. The booming development of sequencing technology and machine learning techniques provide promising alternative approaches for antimicrobial resistance prediction based on sequencing. In this study, we used a lightweight Multitask Learning Transformer to predict the MIC of 14 antibiotics for Salmonella strains based on the genomic information, including point mutations, pan-genome structure, and the profile of antibiotic resistance genes from 5,278 publicly available whole genomes of nontyphoidal Salmonella. And we got better prediction results (improved more than 10% for raw accuracy and 3% for accuracy within ±1 2-fold dilution step) and provided better interpretability than the other ML models. Besides the potential clinical application, our models would cast light on mechanistic understanding of key genetic regions influencing AMR.

****

Bootstrap Motion Forecasting With Self-Consistent Constraints

Maosheng Ye,jiamiao xu,xunnong xu,Tengfei Wang,Tongyi Cao,Qifeng Chen

We present a novel framework to bootstrap Motion forecasting with self-consistent constraints (MISC). The motion forecasting task aims at predicting future trajectories of vehicles by incorporating spatial and temporal information from the past. A key design of MISC is the proposed Dual Consistency Constraints that regularize the predicted trajectories under spatial and temporal perturbation during training. Also, to model the multi-modality in motion forecasting, we design a novel self-ensembling scheme to obtain accurate teacher targets to enforce the self-constraints with multi-modality supervision. With explicit constraints from multiple teacher targets, we observe a clear improvement in the prediction performance. Extensive experiments on the Argoverse motion forecasting benchmark sh

ow that MISC significantly outperforms the state-of-the-art methods. As the proposed strategies are general and can be easily incorporated into other motion forecasting approaches, we also demonstrate that our proposed scheme consistently improves the prediction performance of several existing methods.
****

Learning to Split for Automatic Bias Detection

Yujia Bao,Regina Barzilay

Classifiers are biased when trained on biased datasets. As a remedy, we propose Learning to Split (ls), an algorithm for automatic bias detection. Given a dataset with input-label pairs, ls learns to split this dataset so that predictors trained on the training split cannot generalize to the testing split. This performance gap suggests that the testing split is under-represented in the dataset, which is a signal of potential bias. Identifying non-generalizable splits is challenging since we have no annotations about the bias. In this work, we show that the prediction correctness of each example in the testing split can be used as a source of weak supervision: generalization performance will drop if we move examples that are predicted correctly away from the testing split, leaving only those that are mispredicted. ls is task-agnostic and can be applied to any supervised learning problem, ranging from natural language understanding and image classification to molecular property prediction. Empirical results show that ls is able to generate astonishingly challenging splits that correlate with human-identified biases. Moreover, we demonstrate that combining robust learning algorithms (such as group DRO) with splits identified by ls enables automatic de-biasing. Compared to previous state-of-the-art, we substantially improve the worst-group performance (23.4% on average) when the source of biases is unknown during training and validation. Our code is included in the supplemental materials and will be publicly available.
****

Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Experts

Zhitong Gao,Yucong Chen,Chuyu Zhang,Xuming He

Equipping predicted segmentation with calibrated uncertainty is essential for safety-critical applications. In this work, we focus on capturing the data-inherent uncertainty (aka aleatoric uncertainty) in segmentation, typically when ambiguities exist in input images. Due to the high-dimensional output space and potential multiple modes in segmenting ambiguous images, it remains challenging to predict well-calibrated uncertainty for segmentation. To tackle this problem, we propose a novel mixture of stochastic experts (MoSE) model, where each expert network estimates a distinct mode of the aleatoric uncertainty and a gating network predicts the probabilities of an input image being segmented in those modes. This yields an efficient two-level uncertainty representation. To learn the model, we develop a Wasserstein-like loss that directly minimizes the distribution distance between the MoSE and ground truth annotations. The loss can easily integrate traditional segmentation quality measures and be efficiently optimized via constraint relaxation. We validate our method on the LIDC-IDRI dataset and a modified multimodal Cityscapes dataset. Results demonstrate that our method achieves the state-of-the-art or competitive performance on all metrics.
****

On the Robustness of Safe Reinforcement Learning under Observational Perturbations

Zuxin Liu,Zijian Guo,Zhepeng Cen,Huan Zhang,Jie Tan,Bo Li,Ding Zhao

Safe reinforcement learning (RL) trains a policy to maximize the task reward while satisfying safety constraints. While prior works focus on the performance optimality, we find that the optimal solutions of many safe RL problems are not robust and safe against carefully designed observational perturbations. We formally analyze the unique properties of designing effective observational adversarial attackers in the safe RL setting. We show that baseline adversarial attack techniques for standard RL tasks are not always effective for safe RL and propose two new approaches - one maximizes the cost and the other maximizes the reward. One interesting and counter-intuitive finding is that the maximum reward attack i s strong, as it can both induce unsafe behaviors and make the attack stealthy by maintaining the reward. We further propose a robust training framework for safe RL and evaluate it via comprehensive experiments. This paper provides a pioneer work to investigate the safety and robustness of RL under observational attacks for future safe RL studies. Code is available at: \url{https://github.com/liuzuxin/safe-rl-robustness}

****

Behind the Scenes of Gradient Descent: A Trajectory Analysis via Basis Function Decomposition

Jianhao Ma,Lingjun Guo,Salar Fattahi

This work analyzes the solution trajectory of gradient-based algorithms via a novel basis function decomposition. We show that, although solution trajectories of gradient-based algorithms may vary depending on the learning task, they behave almost monotonically when projected onto an appropriate orthonormal function basis. Such projection gives rise to a basis function decomposition of the solution trajectory. Theoretically, we use our proposed basis function decomposition to establish the convergence of gradient descent (GD) on several representative learning tasks. In particular, we improve the convergence of GD on symmetric matrix factorization and provide a completely new convergence result for the orthogonal symmetric tensor decomposition. Empirically, we illustrate the promise of our proposed framework on realistic deep neural networks (DNNs) across different architectures, gradient-based solvers, and datasets. Our key finding is that gradient-based algorithms monotonically learn the coefficients of a particular orthonormal function basis of DNNs defined as the eigenvectors of the conjugate kernel after training.

****

What Is Missing in IRM Training and Evaluation? Challenges and Solutions

Yihua Zhang,Pranay Sharma,Parikshit Ram,Mingyi Hong,Kush R. Varshney,Sijia Liu

Invariant risk minimization (IRM) has received increasing attention as a way to acquire environment-agnostic data representations and predictions, and also a principled solution for preventing spurious correlations from being learned and improving models’ out-of-distribution generalization. Yet, recent works have found that the optimality of the originally-proposed IRM optimization (IRMV1) may be compromised in practice or could be impossible to achieve in some scenarios. Therefore, a series of advanced IRM algorithms have been developed that show practical improvement over IRMV1. In this work, we revisit these recent IRM advancements and identify and resolve three practical limitations in IRM training and evaluation. First, we find that the effect of batch size during training has been chronically overlooked in previous studies, leaving room for further improvement. We propose small-batch training and highlight the improvements over a set of large-batch optimization techniques. Second, we find that improper selection of evaluation environments could give a false sense of invariance for IRM. To alleviate this effect, we leverage diversified test-time environments to precisely characterize the invariance of IRM when applied in practice. Third, we revisit Ahuja et al. (2020)’s proposal to convert IRM into an ensemble game and identify a limitation when a single invariant predictor is desired instead of an ensemble of individual predictors. We propose a new IRM variant to address this limitation based on a novel viewpoint of ensemble IRM games as consensus-constrained bi-level optimization. Lastly, we conduct extensive experiments (covering 7 existing IRM variants and 7 datasets) to justify the practical significance of revisiting IRM training and evaluation in a principled manner.

****

Neural Decoding of Visual Imagery via Hierarchical Variational Autoencoders

Eleni Miliotou,Panagiotis Kyriakis,Jason D Hinman,Andrei Irimia,Paul Bogdan

Reconstructing natural images from fMRI recordings is a challenging task of great importance in neuroscience. The current architectures are bottlenecked because they fail to effectively capture the hierarchical processing of visual stimuli that takes place in the human brain. Motivated by that fact, we introduce a novel neural network architecture for the problem of neural decoding. Our architecture uses Hierarchical Variational Autoencoders (HVAEs) to learn meaningful repres entations of natural images and leverages their latent space hierarchy to learn voxel-to-image mappings. By mapping the early stages of the visual pathway to the first set of latent variables and the higher visual cortex areas to the deeper layers in the latent hierarchy, we are able to construct a latent variable neural decoding model that replicates the hierarchical visual information processing. Our model achieves better reconstructions compared to the state of the art and our ablation study indicates that the hierarchical structure of the latent space is responsible for that performance.

****

Cooperative Adversarial Learning via Closed-Loop Transcription
Rui Xiao,Xinyu Zhao

This paper proposes a generative model that implements cooperative adversarial learning via closed-loop transcription. In the generative model training, the encoder and decoder are trained simultaneously, and not only the adversarial process but also a cooperative process is included. In the adversarial process, the encoder plays as a critic to maximize the distance between the original and transcribed images, in which the distance is measured by rate reduction in the feature space; in the cooperative process, the encoder and the decoder cooperatively minimize the distance to improve the transcription quality. Cooperative adversarial learning possesses the concepts and properties of Auto-Encoding and GAN, and it is unique in that the encoder actively controls the training process as it is trained in both learning processes in two different roles. Experiments demonstrate that without regularization techniques, our generative model is robust to net architectures and easy to train, sample-wise reconstruction performs well in terms of sample features, and disentangled visual attributes are well modeled in independent principal components.

****

Multi-task Self-supervised Graph Neural Networks Enable Stronger Task Generalization
Mingxuan Ju,Tong Zhao,Qianlong Wen,Wenhao Yu,Neil Shah,Yanfang Ye,Chuxu Zhang

Self-supervised learning (SSL) for graph neural networks (GNNs) has attracted increasing attention from the graph machine learning community in recent years, owing to its capability to learn performant node embeddings without costly label information. One weakness of conventional SSL frameworks for GNNs is that they learn through a single philosophy, such as mutual information maximization or generative reconstruction. When applied to various downstream tasks, these frameworks rarely perform equally well for every task, because one philosophy may not span the extensive knowledge required for all tasks. To enhance the task generalization across tasks, as an important first step forward in exploring fundamental graph models, we introduce PARETOGNN, a multi-task SSL framework for node representation learning over graphs. Specifically, PARETOGNN is self-supervised by manifold pretext tasks observing multiple philosophies. To reconcile different philosophies, we explore a multiple-gradient descent algorithm, such that PARETOGNN actively learns from every pretext task while minimizing potential conflicts. We conduct comprehensive experiments over four downstream tasks (i.e., node classification, node clustering, link prediction, and partition prediction), and our proposal achieves the best overall performance across tasks on 11 widely adopted benchmark datasets. Besides, we observe that learning from multiple philosophies enhances not only the task generalization but also the single task performances, demonstrating that PARETOGNN achieves better task generalization via the disjoint yet complementary knowledge learned from different philosophies. Our code is publicly available at https://github.com/jumxglhf/ParetoGNN.

****

Analyzing Tree Architectures in Ensembles via Neural Tangent Kernel
Ryuichi Kanoh,Mahito Sugiyama

A soft tree is an actively studied variant of a decision tree that updates splitting rules using the gradient method. Although soft trees can take various architectures, their impact is not theoretically well known. In this paper, we formulate and analyze the Neural Tangent Kernel (NTK) induced by soft tree ensembles for arbitrary tree architectures. This kernel leads to the remarkable finding tha t only the number of leaves at each depth is relevant for the tree architecture in ensemble learning with an infinite number of trees. In other words, if the number of leaves at each depth is fixed, the training behavior in function space and the generalization performance are exactly the same across different tree architectures, even if they are not isomorphic. We also show that the NTK of asymmetric trees like decision lists does not degenerate when they get infinitely deep. This is in contrast to the perfect binary trees, whose NTK is known to degenerate and leads to worse generalization performance for deeper trees.

****

Correcting Data Distribution Mismatch in Offline Meta-Reinforcement Learning with Few-Shot Online Adaptation

Jianhao Wang,Jin Zhang,Haozhe Jiang,Junyu Zhang,Liwei Wang,Chongjie Zhang

Offline meta-reinforcement learning (offline meta-RL) extracts knowledge from a given dataset of multiple tasks and achieves fast adaptation to new tasks. Recent offline meta-RL methods typically use task-dependent behavior policies (e.g., training RL agents on each individual task) to collect a multi-task dataset and learn an offline meta-policy. However, these methods always require extra information for fast adaptation, such as offline context for testing tasks or oracle reward functions. Offline meta-RL with few-shot online adaptation remains an open problem. In this paper, we first formally characterize a unique challenge under this setting: data distribution mismatch between offline training and online adaptation. This distribution mismatch may lead to unreliable offline policy evaluation and the regular adaptation methods of online meta-RL will suffer. To address this challenge, we introduce a novel mechanism of data distribution correction, which ensures the consistency between offline and online evaluation by filtering out out-of-distribution episodes in online adaptation. As few-shot out-of-distribution episodes usually have lower returns, we propose a Greedy Context-based data distribution Correction approach, called GCC, which greedily infers how to solve new tasks. GCC diversely samples “task hypotheses” from the current posterior belief and selects a greedy hypothesis with the highest return to update the task belief. Our method is the first to provide an effective online adaptation without additional information, and can be combined with off-the-shelf context-based offline meta-training algorithms. Empirical experiments show that GCC achieves state-of-the-art performance on the Meta-World ML1 benchmark compared to baselines with/without offline adaptation.

****

Sharper Rates and Flexible Framework for Nonconvex SGD with Client and Data Sampling

Alexander Tyurin,Lukang Sun,Konstantin Pavlovich Burlachenko,Peter Richtárik

We revisit the classical problem of finding an approximately stationary point of the average of n smooth and possibly nonconvex functions. The optimal complexity of stochastic first-order methods in terms of the number of gradient evaluations of individual functions is $\mathcal{O}\left(n + n^{1/2}\varepsilon^{-1}\right)$, attained by the optimal SGD methods SPIDER (Cong Fang et al., 2018) and PAGE (Zhize Li et al., 2020), for example, where ε is the error tolerance. However, i) the big-$\mathcal{O}$ notation hides crucial dependencies on the smoothness constants associated with the functions, and ii) the rates and theory in these methods assume simplistic sampling mechanisms that do not offer any flexibility. In this work we remedy the situation. First, we generalize the PAGE algorithm so that it can provably work with virtually any (unbiased) sampling mechanism. This is particularly useful in federated learning, as it allows us to construct and better understand the impact of various combinations of client and data sampling strategies. Second, our analysis is sharper as we make explicit use of certain novel inequalities that capture the intricate interplay between the smoothness constants and the sampling procedure. Indeed, our analysis is better even for the simple sampling procedure analyzed in the PAGE paper. However, this already improved bound can be further sharpened by a different sampling scheme which we propose. In summary, we provide the most general and most accurate analysis of optimal SGD in the smooth nonconvex regime. Finally, our theoretical findings are supposed with carefully designed experiments.

****

Universal embodied intelligence: learning from crowd, recognizing the world, and reinforced with experience

Luo Ji,Longfei Ma,Chang Zhou,Fei Wu,Hongxia Yang

The interactive artificial intelligence in the motion control field is an interesting topic, especially when universal knowledge adaptive to multiple task and universal environments is wanted. Although there are increasing efforts on Reinforcement learning (RL) studies with the assistance of transformers, it might subject to the limitation of the offline training pipeline, in which the exploration and generalization ability is prohibited. Motivated by the cognitive and behavioral psychology, such agent should have the ability to learn from others, recognize the world, and practice itself based its own experience. In this study, we propose the framework of Online Decision MetaMorphFormer (ODM) which attempts to achieve the above learning modes, with a unified model architecture to both highlight its own body perception and produce action and observation predictions. ODM can be applied on any arbitrary agent with a multi-joint body, located in different environments, trained with different type of tasks. Large-scale pretrained dataset are used to warmup ODM while the targeted environment continues to reinforce the universal policy. Substantial interactive experiments as well as few-shot and zero-shot tests in unseen environments and never-experienced tasks verify ODM's performance, and generalization ability. Our study shed some lights on research of general artificial intelligence on the embodied and cognitive field studies.

****

Exploring The Role of Mean Teachers in Self-supervised Masked Auto-Encoders

Youngwan Lee,Jeffrey Ryan Willette,Jonghee Kim,Juho Lee,Sung Ju Hwang

Masked image modeling (MIM) has become a popular strategy for self-supervised learning (SSL) of visual representations with Vision Transformers. A representative MIM model, the masked auto-encoder (MAE), randomly masks a subset of image patches and reconstructs the masked patches given the unmasked patches. Concurrently, many recent works in self-supervised learning utilize the student/teacher paradigm which provides the student with an additional target based on the output of a teacher composed of an exponential moving average (EMA) of previous students. Although common, relatively little is known about the dynamics of the interaction between the student and teacher.

Through analysis on a simple linear model, we find that the teacher conditionally removes previous gradient directions based on feature similarities which effectively acts as a conditional momentum regularizer. From this analysis, we present a simple SSL method, the Reconstruction-Consistent Masked Auto-Encoder (RC-MAE) by adding an EMA teacher to MAE. We find that RC-MAE converges faster and requires less memory usage than state-of-the-art self-distillation methods during pre-training, which may provide a way to enhance the practicality of prohibitively expensive self-supervised learning of Vision Transformer models. Additionally, we show that RC-MAE achieves more robustness and better performance compared to MAE on downstream tasks such as ImageNet-1K classification, object detection, and instance segmentation.

****

Multifactor Sequential Disentanglement via Structured Koopman Autoencoders

Nimrod Berman,Ilan Naiman,Omri Azencot

Disentangling complex data to its latent factors of variation is a fundamental task in representation learning. Existing work on sequential disentanglement mostly provides two factor representations, i.e., it separates the data to time-varying and time-invariant factors. In contrast, we consider multifactor disentanglement in which multiple (more than two) semantic disentangled components are generated. Key to our approach is a strong inductive bias where we assume that the underlying dynamics can be represented linearly in the latent space. Under this assumption, it becomes natural to exploit the recently introduced Koopman autoencoder models. However, disentangled representations are not guaranteed in Koopman approaches, and thus we propose a novel spectral loss term which leads to structured Koopman matrices and disentanglement. Overall, we propose a simple and eas y to code new deep model that is fully unsupervised and it supports multifactor disentanglement. We showcase new disentangling abilities such as swapping of individual static factors between characters, and an incremental swap of disentangled factors from the source to the target. Moreover, we evaluate our method extensively on two factor standard benchmark tasks where we significantly improve over competing unsupervised approaches, and we perform competitively in comparison to weakly- and self-supervised state-of-the-art approaches. The code is available at https://github.com/azencot-group/SKD.

****

Sub-Task Decomposition Enables Learning in Sequence to Sequence Tasks
Noam Wies,Yoav Levine,Amnon Shashua

The field of Natural Language Processing (NLP) has experienced a dramatic leap in capabilities with the recent introduction of huge Language Models (LMs). Despite this success, natural language problems that involve several compounded steps are still practically unlearnable, even by the largest LMs. This complies with experimental failures for end-to-end learning of composite problems that were demonstrated in a variety of domains. An effective mitigation is to introduce intermediate supervision for solving sub-tasks of the compounded problem. Recently, several works have demonstrated high gains by taking a straightforward approach for incorporating intermediate supervision in compounded natural language problems: the sequence-to-sequence LM is fed with an augmented input, in which the decomposed tasks' labels are simply concatenated to the original input. In this paper, we prove a positive learning result that motivates these recent efforts. We show that when concatenating intermediate supervision to the input and training a sequence-to-sequence model on this modified input, unlearnable composite problems can become learnable. We show that this is true for any family of tasks which on the one hand, are unlearnable, and on the other hand, can be decomposed into a polynomial number of simple sub-tasks, each of which depends only on $O(1)$ previous sub-task results. Beyond motivating contemporary empirical efforts for incorporating intermediate supervision in sequence-to-sequence language models, our positive theoretical result is the first of its kind in the landscape of results on the benefits of intermediate supervision for neural-network learning: Until now, all theoretical results on the subject are negative, i.e., show cases where learning is impossible without intermediate supervision, while our result is positive, showing that learning is facilitated in the presence of intermediate supervision.

****

T2D: Spatiotemporal Feature Learning Based on Triple 2D Decomposition
Yucheng Zhao,Chong Luo,Chuanxin Tang,Dongdong Chen,Noel C Codella,Lu Yuan,Zheng-Jun Zha

In this paper, we propose triple 2D decomposition (T2D) of a 3D vision Transformer (ViT) for efficient spatiotemporal feature learning. The idea is to divide the input 3D video data into three 2D data planes and use three 2D filters, implemented by 2D ViT, to extract spatial and motion features. Such a design not only effectively reduces the computational complexity of a 3D ViT, but also guides the network to focus on learning correlations among more relevant tokens. Compared with other decomposition methods, the proposed T2D is shown to be more powerful at a similar computational complexity. The CLIP-initialized T2D-B model achieves state-of-the-art top-1 accuracy of 85.0% and 70.5% on Kinetics-400 and Something-Something-v2 datasets, respectively. It also outperforms other methods by a large margin on FineGym (+17.9%) and Diving-48 (+1.3%) datasets. Under the zero-shot setting, the T2D model obtains a 2.5% top-1 accuracy gain over X-CLIP on HMDB-51 dataset. In addition, T2D is a general decomposition method that can be plugged into any ViT structure of any model size. We demonstrate this by building a tiny size of T2D model based on a hierarchical ViT structure named DaViT. The resulting DaViT-T2D-T model achieves 82.0\% and 71.3\% top-1 accuracy with only 91 GFLOPs on Kinectics-400 and Something-Something-v2 datasets, respectively. Source code will be made publicly available.

****

Online Placebos for Class-incremental Learning

Yaoyao Liu,Yingying Li,Bernt Schiele,Qianru Sun

Not forgetting old class knowledge is a key challenge for class-incremental learning (CIL) when the model continuously adapts to new coming classes. A common technique to address this is knowledge distillation (KD) which penalizes prediction inconsistencies between old and new models. Such prediction is made with almost new class data, as old class data is extremely scarce due to the strict memory limitation in CIL. In this paper, we take a deep dive into KD losses and find that “using new class data for KD” not only hinders the model adaption (for learning new classes) but also results in low efficiency for preserving old class knowledge. We address this by “using the placebos of old classes for KD”, where the placebos are chosen from a free image stream, such as Google Images, in an automatical and economical fashion. To this end, we train an online placebo selection policy to quickly evaluate the quality of streaming images (good or bad placebos) and use only good ones for one-time feed-forward computation of KD. We formulate the policy training process as an online Markov Decision Process (MDP), and introduce an online learning algorithm to solve this MDP problem without causing much computation costs. In experiments, we show that our method 1) is surprisingly effective even when there is no class overlap between placebos and original old class data, 2) does not require any additional supervision or memory budget, and 3) significantly outperforms a number of top-performing CIL methods, in particular when using lower memory budgets for old class exemplars, e.g., five exemplars per class. The code is available in the supplementary.

****

Evaluating Long-Term Memory in 3D Mazes

Jurgis Pašukonis,Timothy P Lillicrap,Danijar Hafner

Intelligent agents need to remember salient information to reason in partially-observed environments. For example, agents with a first-person view should remember the positions of relevant objects even if they go out of view. Similarly, to effectively navigate through rooms agents need to remember the floor plan of how rooms are connected. However, most benchmark tasks in reinforcement learning do not test long-term memory in agents, slowing down progress in this important research direction. In this paper, we introduce the Memory Maze, a 3D domain of randomized mazes specifically designed for evaluating long-term memory in agents. Unlike existing benchmarks, Memory Maze measures long-term memory separate from confounding agent abilities and requires the agent to localize itself by integrating information over time. With Memory Maze, we propose an online reinforcement learning benchmark, a diverse offline dataset, and an offline probing evaluation. Recording a human player establishes a strong baseline and verifies the need to build up and retain memories, which is reflected in their gradually increasing rewards within each episode. We find that current algorithms benefit from training with truncated backpropagation through time and succeed on small mazes, but fall short of human performance on the large mazes, leaving room for future algorithmic designs to be evaluated on the Memory Maze.

****

Packed Ensembles for efficient uncertainty estimation

Olivier Laurent,Adrien Lafage,Enzo Tartaglione,Geoffrey Daniel,Jean-marc Martinez,Andrei Bursuc,Gianni Franchi

Deep Ensembles (DE) are a prominent approach for achieving excellent performance on key metrics such as accuracy, calibration, uncertainty estimation, and out-of-distribution detection. However, hardware limitations of real-world systems constrain to smaller ensembles and lower-capacity networks, significantly deteriorating their performance and properties. We introduce Packed-Ensembles (PE), a strategy to design and train lightweight structured ensembles by carefully modulating the dimension of their encoding space. We leverage grouped convolutions to parallelize the ensemble into a single shared backbone and forward pass to improve training and inference speeds. PE is designed to operate within the memory limits of a standard neural network. Our extensive research indicates that PE accurately preserves the properties of DE, such as diversity, and performs equally well in terms of accuracy, calibration, out-of-distribution detection, and robustness to distribution shift. We make our code available at https://github.com/ENST

A-U2IS/torch-uncertainty.
****

Proactive Multi-Camera Collaboration for 3D Human Pose Estimation
Hai Ci,Mickel Liu,Xuehai Pan,fangwei zhong,Yizhou Wang

This paper presents a multi-agent reinforcement learning (MARL) scheme for proactive Multi-Camera Collaboration in 3D Human Pose Estimation in dynamic human crowds. Traditional fixed-viewpoint multi-camera solutions for human motion capture (MoCap) are limited in capture space and susceptible to dynamic occlusions. Active camera approaches proactively control camera poses to find optimal viewpoints for 3D reconstruction. However, current methods still face challenges with credit assignment and environment dynamics. To address these issues, our proposed method introduces a novel Collaborative Triangulation Contribution Reward (CTCR) that improves convergence and alleviates multi-agent credit assignment issues resulting from using 3D reconstruction accuracy as the shared reward. Additionally, we jointly train our model with multiple world dynamics learning tasks to better capture environment dynamics and encourage anticipatory behaviors for occlusion avoidance. We evaluate our proposed method in four photo-realistic UE4 environments to ensure validity and generalizability. Empirical results show that our method outperforms fixed and active baselines in various scenarios with different numbers of cameras and humans.
****

OpenFE: Automated Feature Generation beyond Expert-level Performance
Tianping Zhang,Zheyu Zhang,Haoyan Luo,Fengyuan Liu,Wei Cao,Jian Li

The goal of automated feature generation is to liberate machine learning experts from the laborious task of manual feature generation, which is crucial for improving the learning performance of tabular data. The major challenge in automated feature generation is to efficiently and accurately identify useful features from a vast pool of candidate features. In this paper, we present OpenFE, an automated feature generation tool that provides competitive results against machine learning experts. OpenFE achieves efficiency and accuracy with two components: 1) a novel feature boosting method for accurately estimating the incremental performance of candidate features. 2) a feature-scoring framework for retrieving effective features from a large number of candidates through successive featurewise halving and feature importance attribution. Extensive experiments on seven benchmark datasets show that OpenFE outperforms existing baseline methods. We further evaluate OpenFE in two famous Kaggle competitions with thousands of data science teams participating. In one of the competitions, features generated by OpenFE with a simple baseline model can beat 99.3% data science teams, demonstrating for the first time that automated feature generation can outperform human experts. In addition to the empirical results, we provide a theoretical perspective to show that feature generation has benefit provably in a simple yet representative setting. Codes and datasets are available in the supplementary materials.
****

Physics-empowered Molecular Representation Learning
Seunghoon Yi,Youngwoo Cho,Jinhwan Sul,Seung Woo Ko,Soo Kyung Kim,Jaegul Choo,Hongkee Yoon,Joonseok Lee

Estimating the energetic properties of molecular systems is a critical task in material design. With the trade-off between accuracy and computational cost, various methods have been used to predict the energy of materials, including recent neural-net-based models. However, most existing neural-net models are context-free (physics-ignoring) black-box models, limiting their applications to predict energy only within the distribution of the training set and thus preventing from being applied to the real practice of molecular design. Inspired by the physical mechanism of the interatomic potential, we propose a physics-driven energy prediction model using a Transformer. Our model is trained not only on the energy regression in the training set, but also with conditions inspired by physical insights and self-supervision based on Masked Atomic Modeling, making it adaptable to the optimization of molecular structure beyond the range observed during training, taking a step towards realizable molecular structure optimization.
****

Revisiting Domain Randomization Via Relaxed State-Adversarial Policy Optimization

Yun-Hsuan Lien,Ping-Chun Hsieh,Yu-Shuen Wang

Domain randomization (DR) is widely used in reinforcement learning (RL) to bridge the gap between simulation and reality through maximizing its average returns under the perturbation of environmental parameters. Although effective, the methods have two limitations: (1) Even the most complex simulators cannot capture all details in reality due to finite domain parameters and simplified physical models. (2) Previous methods often assume that the distribution of domain parameters is a specific family of probability functions, such as a normal or a uniform distribution, which may not be correct. To enable robust RL via DR without the aforementioned limitations, we rethink DR from the perspective of adversarial state perturbation, without the need for re-configuring the simulator or relying on prior knowledge about the environment. We point out that perturbing agents to the worst states during training is naive and could make the agents over-conservative. Hence, we present a Relaxed State-Adversarial Algorithm to tackle the over-conservatism issue by simultaneously maximizing the average-case and worst-case performance of policies. We compared our method to the state-of-the-art methods for evaluation. Experimental results and theoretical proofs verified the effectiveness of our method.

****

Become a Proficient Player with Limited Data through Watching Pure Videos

Weirui Ye,Yunsheng Zhang,Pieter Abbeel,Yang Gao

Recently, RL has shown its strong ability for visually complex tasks. However, it suffers from the low sample efficiency and poor generalization ability, which prevent RL from being useful in real-world scenarios. Inspired by the huge success of unsupervised pre-training methods on language and vision domains, we propose to improve the sample efficiency via a novel pre-training method for model-based RL.

Instead of using pre-recorded agent trajectories that come with their own actions, we consider the setting where the pre-training data are action-free videos, which are more common and available in the real world. We introduce a two-phase training pipeline as follows: for the pre-training phase, we implicitly extract the hidden action embedding from videos and pre-train the visual representation and the environment dynamics network through a novel \Changes{forward-inverse} cycle consistency \Changes{(FICC)} objective based on vector quantization; for down-stream tasks, we finetune with small amount of task data based on the learned models. Our framework can significantly improve the sample efficiency on Atari Games with data of only one hour of game playing. We achieve 118.4\% mean human performance and 36.0\% median performance with only 50k environment steps, which is 85.6\% and 65.1\% better than the scratch EfficientZero model. We believe such pre-training approach can provide an option for solving real-world RL problems. The code is available at \url{https://github.com/YeWR/FICC.git}.

****

Human MotionFormer: Transferring Human Motions with Vision Transformers

Hongyu Liu,Xintong Han,Chenbin Jin,Lihui Qian,Huawei Wei,Zhe Lin,Faqiang Wang,Haoye Dong,Yibing Song,Jia Xu,Qifeng Chen

Human motion transfer aims to transfer motions from a target dynamic person to a source static one for motion synthesis. An accurate matching between the source person and the target motion in both large and subtle motion changes is vital for improving the transferred motion quality. In this paper, we propose Human MotionFormer, a hierarchical ViT framework that leverages global and local perceptions to capture large and subtle motion matching, respectively. It consists of two ViT encoders to extract input features (i.e., a target motion image and a source human image) and a ViT decoder with several cascaded blocks for feature matching and motion transfer. In each block, we set the target motion feature as Query and the source person as Key and Value, calculating the cross-attention maps to conduct a global feature matching. Further, we introduce a convolutional layer to improve the local perception after the global cross-attention computations. This matching process is implemented in both warping and generation branches to guide the motion transfer. During training, we propose a mutual learning loss to enable the co-supervision between warping and generation branches for better motion representations. Experiments show that our Human MotionFormer sets the new state-of-the-art performance both qualitatively and quantitatively. Project page: https://github.com/KumapowerLIU/Human-MotionFormer.

****

Entity Divider with Language Grounding in Multi-Agent Reinforcement Learning

Ziluo Ding,Wanpeng Zhang,Junpeng Yue,Xiangjun Wang,Tiejun Huang,Zongqing Lu

We investigate the use of natural language to drive the generalization of policies in multi-agent settings. Unlike single-agent settings, the generalization of policies should also consider the influence of other agents. Besides, with the increasing number of entities in multi-agent settings, more agent-entity interactions are needed for language grounding, and the enormous search space could impede the learning process. Moreover, given a simple general instruction, e.g., beating all enemies, agents are required to decompose it into multiple subgoals and figure out the right one to focus on. Inspired by previous work, we try to address these issues at the entity level and propose a novel framework for language grounding in multi-agent reinforcement learning, entity divider (EnDi). EnDi enables agents to independently learn subgoal division at the entity level and act in the environment based on the associated entities. The subgoal division is regularized by opponent modeling to avoid subgoal conflicts and promote coordinated strategies. Empirically, EnDi demonstrates the strong generalization ability to unseen games with new dynamics and expresses the superiority over existing methods.

****

Multi-Agent Sequential Decision-Making via Communication

Ziluo Ding,Kefan Su,Weixin Hong,Liwen Zhu,Tiejun Huang,Zongqing Lu

Communication helps agents to obtain information about others so that better coordinated behavior can be learned. Some existing work communicates predicted future trajectory with others, hoping to get clues about what others would do for better coordination. However, circular dependencies sometimes can occur when agents are treated synchronously so it is hard to coordinate decision-making. In this paper, we propose a novel communication scheme, Sequential Communication (SeqComm). SeqComm treats agents asynchronously (the upper-level agents make decisions before the lower-level ones) and has two communication phases. In negotiation phase, agents determine the priority of decision-making by communicating hidden states of observations and comparing the value of intention, which is obtained by modeling the environment dynamics. In launching phase, the upper-level agents take the lead in making decisions and communicate their actions with the lower-level agents. Theoretically, we prove the policies learned by SeqComm are guaranteed to improve monotonically and converge. Empirically, we show that SeqComm outperforms existing methods in various multi-agent cooperative tasks.

****

Hierarchies of Reward Machines

Daniel Furelos-Blanco,Mark Law,Anders Jonsson,Krysia Broda,Alessandra Russo

Reward machines (RMs) are a recent formalism for representing the reward function of a reinforcement learning task through a finite-state machine whose edges encode landmarks of the task using high-level events. The structure of RMs enables the decomposition of a task into simpler and independently solvable subtasks that help tackle long-horizon and/or sparse reward tasks. We propose a formalism for further abstracting the subtask structure by endowing an RM with the ability to call other RMs, thus composing a hierarchy of RMs (HRM). We exploit HRMs by treating each call to an RM as an independently solvable subtask using the options framework, and describe a curriculum-based method to learn HRMs from traces observed by the agent. Our experiments reveal that exploiting a handcrafted HRM leads to faster convergence than with a flat HRM, and that learning an HRM remains feasible in cases where its equivalent flat representation is not.

****

EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conver

sion
Chenfeng Miao,Qingying Zhu,Minchuan Chen,Jun Ma,Shaojun Wang,Jing Xiao

Text-to-speech (TTS) field is recently dominated by one-stage text-to-waveform models, in which the speech quality is significantly improved compared to two-stage models. However, the best-performing open-sourced one-stage model, the VITS, is not fully differentiable and suffers from relatively high computation costs. To address these issues, we propose EfficientTTS 2 (EFTS2), a fully differentiable end-to-end TTS framework that is highly efficient. Our method adopts an adversarial training process, with a differentiable aligner and a hierarchical-VAE-based waveform generator. The differentiable aligner is built upon the EfficientTTS. A hybrid attention mechanism and a variational alignment predictor are incorporated into our network to improve the expressiveness of the aligner. The use of the hierarchical-VAE-based waveform generator not only alleviates the one-to-many mapping problem in waveform generation but also allows the model to learn hierarchical and explainable latent variables that control different aspects of the generated speech. We also extend EFTS2 to the voice conversion (VC) task and propose EFTS2-VC, an end-to-end VC model that allows efficient and high-quality conversion. Experimental results suggest that the two proposed models match their strong counterparts in speech quality with a faster inference speed and smaller model size.

****

LatentAugment: Dynamically Optimized Latent Probabilities of Data Augmentation
Koichi Kuriyama

Although data augmentation is a powerful technique for improving the performance of image classification tasks, it is difficult to identify the best augmentation policy. The optimal augmentation policy, which is the latent variable, cannot be directly observed. To address this problem, this study proposes \textit{LatentAugment}, which estimates the latent probability of optimal augmentation. The proposed method is appealing in that it can dynamically optimize the augmentation strategies for each input and model parameter in learning iterations. Theoretical analysis shows that LatentAugment is a general model that includes other augmentation methods as special cases, and it is simple and computationally efficient in comparison with existing augmentation methods. Experimental results show that the proposed LatentAugment has higher test accuracy than previous augmentation methods on the CIFAR-10, CIFAR-100, SVHN, and ImageNet datasets.

****

Novel Class Discovery under Unreliable Sampling
Haoang Chi,Wenjing Yang,Feng Liu,Long Lan,Tongliang Liu,Tao Qin,Bo Han

When sampling data of specific classes (i.e., known classes) for a scientific task, collectors may encounter unknown classes (i.e., novel classes). Since these novel classes might be valuable for future research, collectors will also sample them and assign them to several clusters with the help of known-class data. This assigning process is also known as novel class discovery (NCD). However, sampling errors are common in practice and may make the NCD process unreliable. To tackle this problem, this paper introduces a new and more realistic setting, where collectors may misidentify known classes and even confuse known classes with novel classes - we name it NCD under unreliable sampling (NUSA). We find that NUSA will empirically degrade existing NCD methods if taking no care of sampling errors. To handle NUSA, we propose an effective solution, named hidden-prototype-based discovery network (HPDN). HPDN first trains a deep network to fully fit the wrongly sampled data, then applies the relatively clean hidden representations yielded by this network into a novel mini-batch K-means algorithm, which further prevents them overfitting to residual errors by detaching noisy supervision timely. Experiments demonstrate that, under NUSA, HPDN significantly outperforms competitive baselines (e.g., 6% more than the best baseline on CIFAR-10) and keeps robust even encountering serious sampling errors.

****

NEW TRAINING FRAMEWORK FOR SPEECH ENHANCEMENT USING REAL NOISY SPEECH

Szu-Wei Fu,Cheng Yu,Yu Tsao,Vishak Gopal,Jayant Gupchup,Ross Cutler

Recently, deep learning-based speech enhancement (SE) models have gained significant improvements. However, the success is mainly based on using synthetic training data created by adding clean speech with noise. On the other hand, in spite of its large amount, real noisy speech is hard to be applied for SE model training because of lack of its clean reference. In this paper, we propose a novel method to utilize real noisy speech for SE model training based on a non-intrusive speech quality prediction model. The SE model is trained through the guide of the quality prediction model. We also find that a speech quality predictor with better accuracy may not necessarily be an appropriate teacher to guide the SE model. In addition, we show that if the quality prediction model is adversarially robust, then the prediction model itself can also be served as a SE model by modifying the input noisy speech through gradient backpropagation. Objective experiment results show that, under the same SE model structure, the proposed new training method trained on a large amount of real noisy speech can outperform the conventional supervised model trained on synthetic noisy speech. Lastly, the two training methods can be combined to utilize both benefits of synthetic noisy speech (easy to learn) and real noisy speech (large amount) to form semi-supervised learning which can further boost the performance both objectively and subjectively. The code will be released after publication.

****

PA-LoFTR: Local Feature Matching with 3D Position-Aware Transformer

Chenhao Li,Qisheng Tang,Shuangjiu Xiao,Ying Mao,Deli Dong,Zhifeng Shi,Guoliang Chen,Jiawen Cheng

We propose a novel image feature matching method that utilizes 3D position information to augment feature representation with a deep neural network. The proposed method introduces 3D position embedding to a state-of-the-art feature matcher, LoFTR, and achieves more promising performance. Following the coarse-to-fine matching pipeline of LoFTR, we construct a Transformer-based neural network that generates dense pixel-wise matches. Instead of using 2D position embeddings for transformer, the proposed method generates 3D position embeddings that can precisely capture position correspondence of matches between images. Importantly, in order to guide neural network to learn 3D space information, we augment features with depth information generated by a depth predictor. In this way, our method, PA-LoFTR, can generate 3D position-aware local feature descriptors with Transformer. We experiment on indoor datasets, and results show that PA-LoFTR improves the performance of feature matching compared to state-of-the-art methods.

****

Policy Contrastive Imitation Learning

Jialei Huang,Zhao-Heng Yin,Yingdong Hu,Yang Gao

Adversarial imitation learning (AIL) is a popular method that has recently achieved much success. However, the performance of AIL is still unsatisfactory on the more challenging tasks. We find that one of the major reasons is due to the low quality of AIL discriminator representation. Since the AIL discriminator is trained via binary classification that does not necessarily discriminate the policy from the expert in a meaningful way, the resulting reward might not be meaningful either. We propose a new method called Policy Contrastive Imitation Learning (PCIL) to resolve this issue. PCIL learns a contrastive representation space by anchoring on different policies and uses a smooth cosine-similarity-based reward to encourage imitation learning. Our proposed representation learning objective can be viewed as a stronger version of the AIL objective and provide a more meaningful comparison between the agent and the policy. From a theoretical perspective, we show the validity of our method using the apprenticeship learning framework. Furthermore, our empirical evaluation on the DeepMind Control suite demonstrates that PCIL can achieve state-of-the-art performance. Finally, qualitative results suggest that PCIL builds a smoother and more meaningful representation space for imitation learning.

****

Backstepping Temporal Difference Learning

Han-Dong Lim,Donghwan Lee

Off-policy learning ability is an important feature of reinforcement learning (RL) for practical applications. However, even one of the most elementary RL algorithms, temporal-difference (TD) learning, is known to suffer form divergence issue when the off-policy scheme is used together with linear function approximation. To overcome the divergent behavior, several off-policy TD learning algorithms have been developed until now. In this work, we provide a unified view of such algorithms from a purely control-theoretic perspective. Our method relies on the backstepping technique, which is widely used in nonlinear control theory.

****

Hidden Markov Transformer for Simultaneous Machine Translation

Shaolei Zhang,Yang Feng

Simultaneous machine translation (SiMT) outputs the target sequence while receiving the source sequence, and hence learning when to start translating each target token is the core challenge for SiMT task. However, it is non-trivial to learn the optimal moment among many possible moments of starting translating, as the moments of starting translating always hide inside the model and can only be supervised with the observed target sequence. In this paper, we propose a Hidden Markov Transformer (HMT), which treats the moments of starting translating as hidden events and the target sequence as the corresponding observed events, thereby organizing them as a hidden Markov model. HMT explicitly models multiple moments of starting translating as the candidate hidden events, and then selects one to generate the target token. During training, by maximizing the marginal likelihood of the target sequence over multiple moments of starting translating, HMT learns to start translating at the moments that target tokens can be generated more accurately. Experiments on multiple SiMT benchmarks show that HMT outperforms strong baselines and achieves state-of-the-art performance.

****

A General Rank Preserving Framework for Asymmetric Image Retrieval

Hui Wu,Min Wang,Wengang Zhou,Houqiang Li

Asymmetric image retrieval aims to deploy compatible models on platforms of different resources to achieve a balance between computational efficiency and retrieval accuracy. The most critical issue is how to align the output features of different models. Despite the great progress, existing approaches apply strong constraints so that features or neighbor structures are strictly aligned across different models. However, such a one-to-one constraint is too strict to be well preserved for the query models with low capacity. Considering that the primary concern of the users is the rank of the returned images, we propose a generic rank preserving framework, which achieves feature compatibility and the order consistency between query and gallery models simultaneously. Specifically, we propose two alternatives to instantiate the framework. One realizes straightforward rank order preservation by directly preserving the consistency of the sorting results. To make sorting process differentiable, the Heaviside step function in sorting is approximated by the sigmoid function. The other aims to preserve a learnable monotonic mapping relationship between the returned similarity scores of query and gallery models. The mapped similarity scores of gallery model are considered as pseudo-supervision to guide the query model training. Extensive experiments on various large-scale datasets demonstrate the superiority of our two proposed methods.

****

Single-level Adversarial Data Synthesis based on Neural Tangent Kernels
Yu-Rong Zhang,Reddy Su,Sheng-Yen Chou,Shan-Hung Wu

Generative adversarial networks (GANs) have achieved impressive performance in data synthesis and have driven the development of many applications. How- ever, GANs are known to be hard to train due to their bilevel objective, which leads to the problems of convergence, mode collapse, and gradient vanishing. In this paper, we propose a new generative model called the generative adversarial NTK (GA-NTK) that has a single-level objective. The GA-NTK keeps the spirit of adversarial learning (which helps generate plausible data) while avoiding the training difficulties of GANs. This is done by modeling the discriminator as a Gaussian process with a neural tangent kernel (NTK-GP) whose training dynam- ics can be completely described by a closed-form formula. We analyze the conver- gence behavior of GA-NTK trained by gradient descent and give some sufficient conditions for convergence. We also conduct extensive experiments to study the advantages and limitations of GA-NTK and propose some techniques that make GA-NTK more practical.

****

Learning to Count Everything: Transformer-based Trackers are Strong Baselines for Class Agnostic Counting
Tsung-Han Chou,Brian Schweitzer Wang,Han Ru Chen,Jun-Cheng Chen

Class agnostic counting (CAC) is a vision task which can be used to count the total occurrence number of any given reference objects in the query image. The task is usually formulated as density map estimation problem through similarity computation among few image samples of the reference object and the query image. In this paper, we show the the popular and effective similarity computation operation, bilinear similarity, actually share high resemblance with self-attention and cross-attention operations which are widely used in the transformer architecture. Inspired by this observation, since the formulation of visual object tracking task is similar to CAC, we show the advanced attention modules of transformer-based trackers are actually powerful matching tools for the CAC task. These modules allow to learn more distinct features to capture the shared patterns among the query and reference images. In addition, we propose a transformer-based class agnostic counting framework by adapting transformer-based trackers for CAC. We demonstrate the effectiveness of the proposed framework with two state-of-the-art transformer-based trackers, MixFormer and TransT, with extensive experiments and ablation studies. The proposed methods outperform other state-of-the-art methods on the challenging FSC-147 and CARPK datasets and achieve new state-of-the-art performances. The code will be publicly available upon acceptance.

****

Unified Algorithms for RL with Decision-Estimation Coefficients: No-Regret, PAC, and Reward-Free Learning
Fan Chen,Song Mei,Yu Bai

Finding unified complexity measures and algorithms for sample-efficient learning is a central topic of research in reinforcement learning (RL). The Decision-Estimation Coefficient (DEC) is recently proposed by Foster et al. (2021) as a necessary and sufficient complexity measure for sample-efficient no-regret RL. This paper makes progress towards a unified theory for RL with the DEC framework. First, we propose two new DEC-type complexity measures: Explorative DEC (EDEC), and Reward-Free DEC (RFDEC). We show that they are necessary and sufficient for sample-efficient PAC learning and reward-free learning, thereby extending the original DEC which only captures no-regret learning. Next, we design new unified sample-efficient algorithms for all three learning goals. Our algorithms instantiate variants of the Estimation-To-Decisions (E2D) meta-algorithm with a strong and general model estimation subroutine. Even in the no-regret setting, our algorithm \textsc{E2D-TA} improves upon the algorithms of Foster et al. (2021) which require either bounding a variant of the DEC which may be prohibitively large, or designing problem-specific estimation subroutines. As applications, we recover existing and obtain new sample-efficient learning results for a wide range of tractable RL problems using essentially a single algorithm. Finally, as a connection, we re-analyze two existing optimistic model-based algorithms based on Posterior Sampling or Maximum Likelihood Estimation, showing that they enjoy similar reg

ret bounds as \textsc{E2D-TA} under similar structural conditions as the DEC.
****
Strength-Adaptive Adversarial Training
Chaojian Yu,Dawei Zhou,Li Shen,Jun Yu,Bo Han,Mingming Gong,Nannan Wang,Tongliang Liu
Adversarial training (AT) is proved to reliably improve network's robustness against adversarial data. However, current AT with a pre-specified perturbation budget has limitations in learning a robust network. Firstly, applying a pre-specified perturbation budget on networks of various model capacities will yield divergent degree of robustness disparity between natural and robust accuracies, which deviates from robust network's desideratum. Secondly, the attack strength of adversarial training data constrained by the pre-specified perturbation budget fails to upgrade as the growth of network robustness, which leads to robust overfitting and further degrades the adversarial robustness. To overcome these limitations, we propose Strength-Adaptive Adversarial Training (SAAT). Specifically, the adversary employs an adversarial loss constraint to generate adversarial training data. Under this constraint, the perturbation budget will be adaptively adjusted according to the training state of adversarial data, which can effectively avoid robust overfitting. Besides, SAAT explicitly constrains the attack strength of training data through the adversarial loss, which manipulates model capacity scheduling during training, and thereby can flexibly control the degree of robustness disparity and adjust the tradeoff between natural accuracy and robustness. Extensive experiments show that our proposal boosts the robustness of adversarial training.
****
Teach me how to Interpolate a Myriad of Embeddings
Shashanka Venkataramanan,Ewa Kijak,laurent amsaleg,Yannis Avrithis
Mixup refers to interpolation-based data augmentation, originally motivated as a way to go beyond empirical risk minimization (ERM). Yet, its extensions focus on the definition of interpolation and the space where it takes place, while the augmentation itself is less studied: For a mini-batch of size m, most methods interpolate between m pairs with a single scalar interpolation factor λ.

In this work, we make progress in this direction by introducing MultiMix, which interpolates an arbitrary number n of tuples, each of length m, with one vector λ per tuple. On sequence data, we further extend to dense interpolation and loss computation over all spatial positions. Overall, we increase the number of tuples per mini-batch by orders of magnitude at little additional cost. This is possible by interpolating at the very last layer before the classifier. Finally, to address inconsistencies due to linear target interpolation, we introduce a self-distillation approach to generate and interpolate synthetic targets.

We empirically show that our contributions result in significant improvement over state-of-the-art mixup methods on four benchmarks. By analyzing the embedding space, we observe that the classes are more tightly clustered and uniformly spread over the embedding space, thereby explaining the improved behavior.
****
Mega: Moving Average Equipped Gated Attention
Xuezhe Ma,Chunting Zhou,Xiang Kong,Junxian He,Liangke Gui,Graham Neubig,Jonathan May,Luke Zettlemoyer
The design choices in the Transformer attention mechanism, including weak inductive bias and quadratic computational complexity, have limited its application for modeling long sequences. In this paper, we introduce Mega, a simple, theoretically grounded, single-head gated attention mechanism equipped with (exponential) moving average to incorporate inductive bias of position-aware local dependencies into the position-agnostic attention mechanism. We further propose a variant of Mega that offers linear time and space complexity yet yields only minimal quality loss, by efficiently splitting the whole sequence into multiple chunks with fixed length. Extensive experiments on a wide range of sequence modeling bench

marks, including the Long Range Arena, neural machine translation, auto-regressive language modeling, and image and speech classification, show that Mega achieves significant improvements over other sequence models, including variants of Transformers and recent state space models.

****

IEDR: A Context-aware Intrinsic and Extrinsic Disentangled Recommender System

Yixin Su,Wei Jiang,Yunxiang Zhao,Fangquan Lin,Cheng Yang,Sarah Monazam Erfani,Junhao Gan

Intrinsic and extrinsic factors jointly affect users' decisions in item selection (e.g., click, purchase). Intrinsic factors reveal users' real interests and are invariant in different contexts (e.g., time, weather), whereas extrinsic factors can change w.r.t. different contexts. Analyzing these two factors is an essential yet challenging task in recommender systems. However, in existing studies, factor analysis is either largely neglected, or designed for a specific context (e.g., the time context in sequential recommendation), which limits the applicability of such models. In this paper, we propose a generic model, IEDR, to learn intrinsic and extrinsic factors from various contexts for recommendation. IEDR contains two key components: a contrastive learning component, and a disentangling component. The two components collaboratively enable our model to learn context-invariant intrinsic factors and context-based extrinsic factors from all available contexts. Experimental results on real-world datasets demonstrate the effectiveness of our model in factor learning and impart a significant improvement in recommendation accuracy over the state-of-the-art methods.

****

Deep Deformation Based on Feature-Constraint for 3D Human Mesh Correspondence

Na Gong

In this study, we address the challenges in mesh correspondence for various types of complete or single-view human body data. The parametric human model has been widely used in various human-related applications and in 3D human mesh correspondence because it provides sufficient scope to modify the resulting model. In contrast to prior methods that optimize both the correspondences and human model parameters (pose and shape), some of the recent methods directly deform each vertex of a parametric template by processing the point clouds that represent the input shapes. This allows the models to have more accurate representations of the details while maintaining the correspondence. However, we identified two limitations in these methods. First, it is difficult for the transformed template to completely restore the input shapes using only a pointwise reconstruction loss. Second, they cannot deform the template to a single-view human body from the depth camera observations or infer the correspondences between various forms of input human bodies. In representation learning, one of the main challenges is to design appropriate loss functions for supervising features with different abilities. To address this, we introduce the feature constraint deformation network (FCD-Net), which is an end-to-end deep learning approach that identifies 3D human mesh correspondences by learning various shape transformations from a predetermined template. The FCD-Net is implemented by an encoder–decoder architecture. A global feature encoded from the input shape and a decoder are used to deform the template based on the encoded global feature. We simultaneously input the complete shape and single-view shape into the encoder and closely constrain the features to enable the encoder to learn more robust features. Meanwhile, the decoder generates a completely transformed template with higher promise by using the complete shape as the ground truth, even if the input is single-view human body data. We conduct extensive experiments to validate the effectiveness of the proposed FCD-Net on four types of single-view human body data, both from qualitative and quantitative aspects. We also demonstrate that our approach improves the state-of-the-art results on the difficult "FAUST-inter" and "SHREC'19" challenges, with average correspondence errors of 2.54 cm and 6.62 cm, respectively . In addition, the proposed FCD-Net performs well on real and unclean point clouds from a depth camera.

****

Explaining Representation Bottlenecks of Convolutional Decoder Networks
Ling Tang,Wen Shen,Zhanpeng Zhou,YueFeng Chen,Quanshi Zhang

In this paper, we prove representation bottlenecks of a cascaded convolutional decoder network, considering the capacity of representing different frequency components of an input sample. We conduct the discrete Fourier transform on each channel of the feature map in an intermediate layer of the decoder network. Then, we introduce the rule of the forward propagation of such intermediate-layer spectrum maps, which is equivalent to the forward propagation of feature maps through a convolutional layer. Based on this, we find that each frequency component in the spectrum map is forward propagated independently with other frequency components. Furthermore, we prove two bottlenecks in representing feature spectrums. First, we prove that the convolution operation, the zero-padding operation, and a set of other settings all make a convolutional decoder network more likely to weaken high-frequency components. Second, we prove that the upsampling operation generates a feature spectrum, in which strong signals repetitively appears at certain frequencies. We will release all codes when this paper is accepted.

****

Dual Ensembled Multiagent Q-Learning with Hypernet Regularizer
Yaodong Yang,Guangyong Chen,Hongyao Tang,Furui Liu,Danruo DENG,Jianye HAO,Pheng-Ann Heng

Overestimation in the temporal-difference single-agent reinforcement learning has been widely studied, where the variance in value estimation causes overestimation of the maximal target value due to Jensen's inequality. Instead, overestimation in multiagent settings has received little attention though it can be even more severe. One kind of pioneer work extends ensemble methods from single-agent deep reinforcement learning to address the multiagent overestimation by discarding the large target values among the ensemble. However, its ability is limited by the ensemble diversity. Another kind of work softens the maximum operator in the Bellman equation to avoid large target values, but also leads to sub-optimal value functions. Unlike previous works, in this paper, we address the multiagent overestimation by analyzing its underlying causes in an estimation-optimization iteration manner. We show that the overestimation in multiagent value-mixing Q-learning not only comes from the overestimation of target Q-values but also accumulates in the online Q-network's optimization step. Therefore, first, we integrate the random ensemble and in-target minimization into the estimation of target Q-values to derive a lower update target. Second, we propose a novel hypernet regularizer on the learnable terms of the online global Q-network to further reduce overestimation. Experiments on various kinds of tasks demonstrate that the proposed method consistently addresses the overestimation problem while previous works fail.

****

Exploring Chemical Space with Score-based Out-of-distribution Generation
Seul Lee,Jaehyeong Jo,Sung Ju Hwang

A well-known limitation of existing works on molecule generation is that the generated molecules highly resemble those in the training set. To generate truly novel molecules with completely different structures that may have even better properties than known molecules for de novo drug discovery, more powerful exploration in the chemical space is necessary. To this end, we propose Molecular Out-Of-distribution Diffusion (MOOD), a novel score-based diffusion scheme that incorporates out-of-distribution (OOD) control in the generative stochastic differential equation (SDE) with simple control of a hyperparameter, thus requires no additional computational costs unlike existing methods (e.g., RL-based methods). However, some novel molecules may be chemically implausible, or may not meet the basic requirements of real-world drugs. Thus, MOOD performs conditional generation by utilizing the gradients from a property prediction network that guides the reverse-time diffusion process to high-scoring regions according to multiple target properties such as protein-ligand interactions, drug-likeness, and synthesizability. This allows MOOD to search for novel and meaningful molecules rather than generating unseen yet trivial ones. We experimentally validate that MOOD is able to explore the chemical space beyond the training distribution, generating mol ecules that outscore ones found with existing methods, and even the top 0.01% of the original training pool.

****

Node Number Awareness Representation for Graph Similarity Learning

JaX Lyu,Liang Zhang,Yi Huang,Shifeng Chen,Guangming Zhu,Mohammed Bennamoun,Syed Afaq Ali Shah

This work aims to address two important issues in the graph similarity computation, the first one is the Node Number Awareness Issue (N^{2}AI), and the second one is how to accelerate the inference speed of graph similarity computation in downstream tasks. We found that existing Graph Neural Network based graph similarity models have a large error in predicting the similarity scores of two graphs with similar number of nodes. Our analysis shows that this is because of the global pooling function in graph neural networks that maps graphs with similar number of nodes to similar embedding distributions, reducing the separability of their embeddings, which we refer to as the N^{2}AI. Our motivation is to enhance the difference between the two embeddings to improve their separability, thus we leverage our proposed Different Attention (DiffAtt) to construct Node Number Awareness Graph Similarity Model (N^{2}AGim). In addition, we propose the Graph Similarity Learning with Landmarks (GSL^2) to accelerate similarity computation. GSL^2 uses the trained N^{2}AGim to generate the individual embedding for each graph without any additional learning, and this individual embedding can effectively help GSL^2 to improve its inference speed. Experiments demonstrate that our N^{2}AGim outperforms the second best approach on Mean Square Error by 24.3\%(1.170 vs 1.546), 43.1\%(0.066 vs 0.116), and 44.3\%(0.308 vs 0.553), on AIDS700nef, LINUX, and IMDBMulti datasets, respectively. Our GSL^2 is at most 47.7 and 1.36 times faster than N^{2}AGim and the second faster model. Our code is publicly available on https://github.com/iclr231312/N2AGim.

****

Evaluating Fairness Without Sensitive Attributes: A Framework Using Only Auxiliary Models

Zhaowei Zhu,Yuanshun Yao,Jiankai Sun,Yang Liu,Hang Li

Although the volume of literature and public attention on machine learning fairness has been growing significantly in recent years, in practice some tasks as basic as measuring fairness, which is the first step in studying and promoting fairness, can be challenging. This is because the sensitive attributes are often unavailable in a machine learning system due to privacy regulations. The straightforward solution is to use auxiliary models to predict the missing sensitive attributes. However, our theoretical analyses show that the estimation error of the directly measured fairness metrics is proportional to the error rates of auxiliary models' predictions. Existing works that attempt to reduce the estimation error often require strong assumptions, e.g. access to the ground-truth sensitive attributes in a subset of samples, auxiliary models' training data and the target data are i.i.d, or some form of conditional independence. In this paper, we drop those assumptions and propose a framework that uses only off-the-shelf auxiliary models. The main challenge is how to reduce the negative impact of imperfectly predicted sensitive attributes on the fairness metrics without knowing the ground-truth sensitive attribute values. Inspired by the noisy label learning literature, we first derive a closed-form relationship between the directly measured fairness metrics and their corresponding ground-truth metrics. And then we estimate some key statistics (most importantly transition matrix in the noisy label literature), which we use, together with the derived relationship, to calibrate the fairness metrics. Our framework can be applied to all popular group fairness definitions as well as multi-class classifiers and multi-category sensitive attributes. In addition, we theoretically prove the upper bound of the estimation error in our calibrated metrics and show our method can substantially decrease the estimation error especially when auxiliary models are inaccurate or the target model is highly biased. Experiments on COMPAS and CelebA validate our theoretical analyses and show our method can measure fairness significantly more accurately than baselines under favorable circumstances.

****

Optimal Neural Network Approximation of Wasserstein Gradient Direction via Convex Optimization
Yifei Wang,Peng Chen,Mert Pilanci,Wuchen Li
The computation of Wasserstein gradient direction is essential for posterior sampling problems and scientific computing. The approximation of the Wasserstein gradient with finite samples requires solving a variational problem. We study the variational problem in the family of two-layer networks with squared-ReLU activations, towards which we derive a semi-definite programming (SDP) relaxation. This SDP can be viewed as an approximation of the Wasserstein gradient in a broader function family including two-layer networks. By solving the convex SDP, we obtain the optimal approximation of the Wasserstein gradient direction in this class of functions. Numerical experiments including PDE-constrained Bayesian inference and parameter estimation in COVID-19 modeling demonstrate the effectiveness of the proposed method.
****
Parallel Deep Neural Networks Have Zero Duality Gap
Yifei Wang,Tolga Ergen,Mert Pilanci
Training deep neural networks is a challenging non-convex optimization problem. Recent work has proven that the strong duality holds (which means zero duality gap) for regularized finite-width two-layer ReLU networks and consequently provided an equivalent convex training problem. However, extending this result to deeper networks remains to be an open problem. In this paper, we prove that the duality gap for deeper linear networks with vector outputs is non-zero. In contrast, we show that the zero duality gap can be obtained by stacking standard deep networks in parallel, which we call a parallel architecture, and modifying the regularization. Therefore, we prove the strong duality and existence of equivalent convex problems that enable globally optimal training of deep networks. As a by-product of our analysis, we demonstrate that the weight decay regularization on the network parameters explicitly encourages low-rank solutions via closed-form expressions. In addition, we show that strong duality holds for three-layer standard ReLU networks given rank-1 data matrices.
****
Multi-domain image generation and translation with identifiability guarantees
Shaoan Xie,Lingjing Kong,Mingming Gong,Kun Zhang
Multi-domain image generation and unpaired image-to-to-image translation are two important and related computer vision problems. The common technique for the two tasks is the learning of a joint distribution from multiple marginal distributions. However, it is well known that there can be infinitely many joint distributions that can derive the same marginals. Hence, it is necessary to formulate suitable constraints to address this highly ill-posed problem. Inspired by the recent advances in nonlinear Independent Component Analysis (ICA) theory, we propose a new method to learn the joint distribution from the marginals by enforcing a specific type of minimal change across domains. We report one of the first results connecting multi-domain generative models to identifiability and shows why identifiability is essential and how to achieve it theoretically and practically. We apply our method to five multi-domain image generation and six image-to-image translation tasks. The superior performance of our model supports our theory and demonstrates the effectiveness of our method. The training code are available at https://github.com/Mid-Push/i-stylegan.
****
Interventional Rationalization
Linan Yue,Qi Liu,Li Wang,Yanqing An,Yichao Du,Zhenya Huang
Selective rationalizations improve the explainability of neural networks by selecting a subsequence of the input (i.e., rationales) to explain the prediction results. Although existing methods have achieved promising results, they still suffer from adopting the spurious correlations in data (aka., shortcuts) to compose rationales and make predictions. Inspired by the causal theory, in this paper, we develop an interventional rationalization (Inter-RAT) to discover the causal rationales. Specifically, we first analyse the causalities among the input, rationales and results with a structural causal model. Then, we discover spurious co

rrelations between input and rationales, and between rationales and results, respectively, by identifying the confounder in the causalities. Next, based on the backdoor adjustment, we propose a causal intervention method to remove the spurious correlations in input and rationales. Further, we discuss reasons why spurious correlations between the selected rationales and results exist by analysing the limitations of the sparsity constraint in the rationalization, and employ the causal intervention method to remove these correlations. Extensive experimental results on three real-world datasets clearly validate the effectiveness of our proposed method.

****

Information-Theoretic Analysis of Unsupervised Domain Adaptation

Ziqiao Wang,Yongyi Mao

This paper uses information-theoretic tools to analyze the generalization error in unsupervised domain adaptation (UDA). We present novel upper bounds for two notions of generalization errors. The first notion measures the gap between the population risk in the target domain and that in the source domain, and the second measures the gap between the population risk in the target domain and the empirical risk in the source domain. While our bounds for the first kind of error are in line with the traditional analysis and give similar insights, our bounds on the second kind of error are algorithm-dependent, which also provide insights into algorithm designs. Specifically, we present two simple techniques for improving generalization in UDA and validate them experimentally.

****

Pessimism in the Face of Confounders: Provably Efficient Offline Reinforcement Learning in Partially Observable Markov Decision Processes

Miao Lu,Yifei Min,Zhaoran Wang,Zhuoran Yang

We study offline reinforcement learning (RL) in partially observable Markov decision processes. In particular, we aim to learn an optimal policy from a dataset collected by a behavior policy which possibly depends on the latent state. Such a dataset is confounded in the sense that the latent state simultaneously affects the action and the observation, which is prohibitive for existing offline RL algorithms. To this end, we propose the \underline{P}roxy variable \underline{P}essimistic \underline{P}olicy \underline{O}ptimization (\texttt{P3O}) algorithm, which addresses the confounding bias and the distributional shift between the optimal and behavior policies in the context of general function approximation. At the core of \texttt{P3O} is a coupled sequence of pessimistic confidence regions constructed via proximal causal inference, which is formulated as minimax estimation. Under a partial coverage assumption on the confounded dataset, we prove that \texttt{P3O} achieves a $n^{-1/2}$-suboptimality, where n is the number of trajectories in the dataset. To our best knowledge, \texttt{P3O} is the first provably efficient offline RL algorithm for POMDPs with a confounded dataset.

****

DELVING INTO THE HIERARCHICAL STRUCTURE FOR EFFICIENT LARGE-SCALE BI-LEVEL LEARNING

Xixi Jia,Renzhen Wang,Deyu Meng,Xiangchu Feng

Recent years have witnessed growing interest and emerging successes of bi-level learning in a wide range of applications, such as meta learning and hyper-parameter optimization. While current bi-level learning approaches suffer from high memory and computation costs especially for large-scale deep learning scenarios, which is due to the hierarchical optimization therein. {\textit {It is therefore interesting to know whether the hierarchical structure can be untied for efficient learning}.} To answer this question, we introduce NSGame that, transforming the hierarchical bi-level learning problem into a parallel Nash game, incorporates the tastes of hierarchy by a very small scale Stackelberg game.

We prove that strong differential Stackelberg equilibrium (SDSE) of the bi-level learning problem corresponds to local Nash equilibrium of the NSGame. To obtain such SDSE from NSGame, we introduce a two-time scale stochastic gradient descent (TTS-SGD) method, and provide theoretical guarantee that local Nash equilibrium obtained by the TTS-SGD method is SDSE of the bi-level learning problem. We co mpare NSGame with representative bi-level learning models, such as MWN and MLC, experimental results on class imbalance learning and noisy label learning have verified that the proposed NSGame achieves comparable and even better results than the corresponding meta learning models, while NSGame is computationally more efficient.

****

Can GNNs Learn Heuristic Information for Link Prediction?

Shuming Liang,Yu Ding,Zhidong Li,Bin Liang,siqi zhang,Yang Wang,Fang Chen

Graph Neural Networks (GNNs) have shown superior performance in Link Prediction (LP). Especially, SEAL and its successors address the LP problem by classifying the subgraphs extracted specifically for candidate links, gaining state-of-the-art results. Nevertheless, we question whether these methods can effectively learn the information equivalent to link heuristics such as Common Neighbors, Katz index, etc. (we refer to such information as heuristic information in this work). We show that link heuristics and GNNs capture different information. Link heuristics usually collect pair-specific information by counting the involved neighbors or paths between two nodes in a candidate link, while GNNs learn node-wise representations through a neighborhood aggregation algorithm in which two nodes in the candidate link do not pay special attention to each other. Our further analysis shows that SEAL-type methods only use a GNN to model the pair-specific subgraphs and also cannot effectively capture heuristic information. To verify our analysis, a straightforward way is to compare the LP performance between existing methods and a model that learns heuristic information independently of the GNN learning. To this end, we present a simple yet light framework ComHG by directly Combining the embeddings of link Heuristics and the representations produced by a GNN. Experiments on OGB LP benchmarks show that ComHG outperforms all top competitors by a large margin, empirically confirming our propositions. Our experimental study also indicates that the contributions of link heuristics and the GNN to LP are sensitive to the graph degree, where the former is powerful on sparse graphs while the latter becomes dominant on dense graphs.

****

Understanding Zero-shot Adversarial Robustness for Large-Scale Models

Chengzhi Mao,Scott Geng,Junfeng Yang,Xin Wang,Carl Vondrick

Pretrained large-scale vision-language models like CLIP have exhibited strong generalization over unseen tasks. Yet imperceptible adversarial perturbations can significantly reduce CLIP's performance on new tasks. In this work, we identify and explore the problem of adapting large-scale models for zero-shot adversarial robustness. We first identify two key factors during model adaption--training losses and adaptation methods--that affect the model's zero-shot adversarial robustness. We then propose a text-guided contrastive adversarial training loss, which aligns the text embeddings and the adversarial visual features with contrastive learning on a small set of training data. We apply this training loss to two adaption methods, model finetuning and visual prompt tuning. We find that visual prompt tuning is more effective in the absence of texts, while finetuning wins in the existence of text guidance. Overall, our approach significantly improves the zero-shot adversarial robustness over CLIP, seeing an average improvement of 31 points over ImageNet and 15 zero-shot datasets. We hope this work can shed light on understanding the zero-shot adversarial robustness of large-scale models.

****

HOYER REGULARIZER IS ALL YOU NEED FOR EXTREMELY SPARSE SPIKING NEURAL NETWORKS

Gourav Datta,Zeyu Liu,Peter Anthony Beerel

Spiking Neural networks (SNN) have emerged as an attractive spatio-temporal computing paradigm for a wide range of low-power vision tasks. However, state-of-the-art (SOTA) SNN models either incur multiple time steps which hinder their deployment in real-time use cases or increase the training complexity significantly.

To mitigate this concern, we present a training framework (from scratch) for one-time-step SNNs that uses a novel variant of the recently proposed Hoyer regulari

zer. We estimate the threshold of each SNN layer as the Hoyer extremum of a clipped version of its activation map, where the clipping threshold is trained using
gradient descent with our Hoyer regularizer. This approach not only downscales the value of the trainable threshold, thereby emitting a large number of spikes for
weight update with limited number of iterations (due to only one time step) but also shifts the pre-activation values away from the threshold, thereby mitigating
the effect of noise that can degrade the SNN accuracy. Our approach outperforms existing spiking, binary, and adder neural networks in terms of accuracy-FLOPs trade-off for complex image recognition tasks. Downstream experiments on object detection also demonstrate the efficacy of our approach.
****

Rademacher Complexity Over $\mathcal{H} \Delta \mathcal{H}$ Class for Adversarially Robust Domain Adaptation

Yuyang Deng,Nidham Gazagnadou,Junyuan Hong,Mehrdad Mahdavi,Lingjuan Lyu

In domain adaptation, a model is trained on a dataset generated from a source domain and its generalization is evaluated on a possibly different target domain. Understanding the generalization capability of the learned model is a longstanding question. Recent studies demonstrated that the adversarial robust learning under ℓ_∞ attack is even harder to generalize to different domains. To thoroughly study the fundamental difficulty behind adversarially robust domain adaptation, we propose to analyze a key complexity measure that controls the cross-domain generalization: the adversarial Rademacher complexity over $\mathcal{H} \Delta \mathcal{H}$ class. For linear models, we show that adversarial Rademacher complexity over $\mathcal{H} \Delta \mathcal{H}$ class is always greater than the non-adversarial one, which reveals the intrinsic hardness of adversarially robust domain adaptation. We also establish upper bounds on this complexity measure, and extend them to the ReLU neural network class as well. Finally, by properly extending our generalization bound for adversarially robust domain adaptation, we explain \emph{why adversarial training can help transferring the model performance to different domains}. We believe our results initiate the study of the generalization theory of adversarially robust domain adaptation, and could shed lights on distributed adversarially robust learning from heterogeneous sources -- a scenario typically encountered in federated learning applications.
****

Continual evaluation for lifelong learning: Identifying the stability gap

Matthias De Lange,Gido M van de Ven,Tinne Tuytelaars

Time-dependent data-generating distributions have proven to be difficult for gradient-based training of neural networks, as the greedy updates result in catastrophic forgetting of previously learned knowledge. Despite the progress in the field of continual learning to overcome this forgetting, we show that a set of common state-of-the-art methods still suffers from substantial forgetting upon starting to learn new tasks, except that this forgetting is temporary and followed by a phase of performance recovery. We refer to this intriguing but potentially problematic phenomenon as the stability gap. The stability gap had likely remained under the radar due to standard practice in the field of evaluating continual learning models only after each task. Instead, we establish a framework for continual evaluation that uses per-iteration evaluation and we define a new set of metrics to quantify worst-case performance. Empirically we show that experience replay, constraint-based replay, knowledge-distillation, and parameter regularization methods are all prone to the stability gap; and that the stability gap can be observed in class-, task-, and domain-incremental learning benchmarks. Additionally, a controlled experiment shows that the stability gap increases when tasks are more dissimilar. Finally, by disentangling gradients into plasticity and stability components, we propose a conceptual explanation for the stability gap.
****

On the Universal Approximation Property of Deep Fully Convolutional Neural Networks

Ting Lin,Zuowei Shen,Qianxiao Li

We study the approximation of shift-invariant or equivariant functions by deep fully convolutional networks from the dynamical systems perspective. We prove that deep residual fully convolutional networks and their continuous-layer counterpart can achieve universal approximation of these symmetric functions at constant channel width. Moreover, we show that the same can be achieved by non-residual variants with at least 2 channels in each layer and convolutional kernel size of at least 2. In addition, we show that these requirements are necessary, in the sense that networks with fewer channels or smaller kernels fail to be universal approximators.

****

Can We Faithfully Represent Absence States to Compute Shapley Values on a DNN?

Jie Ren,Zhanpeng Zhou,Qirui Chen,Quanshi Zhang

Masking some input variables of a deep neural network (DNN) and computing output changes on the masked input sample represent a typical way to compute attributions of input variables in the sample. People usually mask an input variable using its baseline value. However, there is no theory to examine whether baseline value faithfully represents the absence of an input variable, i.e., removing all signals from the input variable. Fortunately, recent studies (Ren et al., 2023a; Deng et al., 2022a) show that the inference score of a DNN can be strictly disentangled into a set of causal patterns (or concepts) encoded by the DNN. Therefore, we propose to use causal patterns to examine the faithfulness of baseline values. More crucially, it is proven that causal patterns can be explained as the elementary rationale of the Shapley value. Furthermore, we propose a method to learn optimal baseline values, and experimental results have demonstrated its effectiveness.

****

FedGSNR: Accelerating Federated Learning on Non-IID Data via Maximum Gradient Signal to Noise Ratio

Qi Tan,Yi Zhao,Ke Xu,Qi Li

Federated learning (FL) allows participants jointly training a model without direct data sharing. In such a process, participants rather than the central server perform local updates of stochastic gradient descent (SGD) and the central server aggregates the gradients from the participants to update the global model. However, the non-iid training data in participants significantly impact global model convergence.Most of existing studies addressed this issue by utilizing variance reduction or regularization. However, these studies focusing on specific datasets lack theoretical guarantee for efficient model training. In this paper, we provide a novel perspective on the non-iid issue by optimizing Gradient Signal to Noise Ratio (GSNR) during model training. In each participant, we decompose local gradients calculated on the non-iid training data into the signal and noise components and then speed up the model convergence by maximizing GSNR. We prove that GSNR can be maximized by using the optimal number of local updates. Subsequently, we develop FedGSNR to compute the optimal number of local updates for each participant, which can be applied to existing gradient calculation algorithms to accelerate the global model convergence. Moreover, according to the positive correlation between GSNR and the quality of shared information, FedGSNR allows the server to accurately evaluate contributions of different participants (i.e., the quality of local datasets) by utilizing GSNR. Extensive experimental evaluations demonstrate that FedGSNR achieves on average a 1.69× speedup with comparable accuracy.

****

Dataless Knowledge Fusion by Merging Weights of Language Models

Xisen Jin,Xiang Ren,Daniel Preotiuc-Pietro,Pengxiang Cheng

Fine-tuning pre-trained language models has become the prevalent paradigm for building downstream NLP models. Oftentimes fine-tuned models are readily available but their training data is not, due to data privacy or intellectual property concerns. This creates a barrier to fusing knowledge across individual models to yield a better single model. In this paper, we study the problem of merging individual models built on different training data sets to obtain a single model that

performs well both across all data set domains and can generalize on out-of-domain data. We propose a data-less knowledge fusion method that merges models in their parameter space, guided by weights that minimize prediction differences between the merged model and the individual models. Over a battery of evaluation settings, we show that the proposed method significantly outperforms baselines such as Fisher-weighted averaging or model ensembling. Further, we find that our method is a promising alternative to multi-task learning that can preserve or sometimes improve over the individual models without access to the training data. Finally, model merging is more efficient than training a multi-task model, thus making it applicable to a wider set of scenarios.
****

Domain-Indexing Variational Bayes: Interpretable Domain Index for Domain Adaptation

Zihao Xu,Guang-Yuan Hao,Hao He,Hao Wang

Previous studies have shown that leveraging "domain index" can significantly boost domain adaptation performance (Wang et al., 2020; Xu et al., 2022). However, such domain indices are not always available. To address this challenge, we first provide a formal definition of domain index from the probabilistic perspective, and then propose an adversarial variational Bayesian framework that infers domain indices from multi-domain data, thereby providing additional insight on domain relations and improving domain adaptation performance. Our theoretical analysis shows that our adversarial variational Bayesian framework finds the optimal domain index at equilibrium. Empirical results on both synthetic and real data verify that our model can produce interpretable domain indices which enable us to achieve superior performance compared to state-of-the-art domain adaptation methods. Code is available at https://github.com/Wang-ML-Lab/VDI.
****

Improving the Transferability of Adversarial Attacks through Experienced Precise Nesterov Momentum

Hao Wu,Jinwei Wang,Jiawei Zhang,Bin Ma,Xiangyang Luo,wang yu

Deep Neural Networks are vulnerable to adversarial attacks, which makes adversarial attacks serve as a method to evaluate the robustness of DNNs. However, adversarial attacks have high white-box attack success rates but poor transferability, making black-box attacks impracticable in the real world. Momentum-based attacks were proposed to accelerate optimization to improve transferability. Nevertheless, conventional momentum-based attacks accelerate optimization inefficiently during early iterations since the initial value of momentum is zero, which leads to unsatisfactory transferability. Therefore, we propose Experienced Momentum (EM), which is the pre-trained momentum. Initializing the momentum to EM can help accelerate optimization during the early iterations. Moreover, the pre-update of conventional Nesterov momentum based attacks is rough, prompting us to propose Precise Nesterov momentum (PN). PN refines the pre-update by considering the gradient of the current data point. Finally, we integrate EM with PN as Experienced Precise Nesterov momentum (EPN) to further improve transferability. Extensive experiments against normally trained and defense models demonstrate that our EPN is more effective than conventional momentum in the improvement of transferability. Specifically, the attack success rates of our EPN-based attacks are $\sim$11.9% and $\sim$13.1% higher than conventional momentum-based attacks on average against normally trained and defense models, respectively.
****

TaylorNet: A Taylor-Driven Generic Neural Architecture

Hongjue Zhao,Yizhuo Chen,Dachun Sun,Yingdong Hu,Kaizhao Liang,Yanbing Mao,Lui Sha,Huajie Shao

Physics-informed machine learning (PIML) aims to incorporate physics knowledge into deep neural networks (DNNs) to improve the model generalization. However, existing methods in PIML are either designed for specific problems or hard to interpret the results using black-box DNNs. In this work, we propose Taylor Neural Network (TaylorNet), a generic neural architecture that parameterizes Taylor polynomials using DNNs without using non-linear activation functions. The key challenges of developing TaylorNet lie in: (i) mitigating the curse of dimensionality caused by higher-order terms, and (ii) improving the stability of model training. To overcome these challenges, we first adopt Tucker decomposition to decompose the higher-order derivatives in Taylor expansion parameterized by DNNs into low-rank tensors. Then we propose a novel reducible TaylorNet to further reduce the computational complexity by removing more redundant parameters in the hidden layers. In order to improve training accuracy and stability, we develop a new Taylor initialization method. Finally, the proposed models are evaluated on a broad spectrum of applications, including image classification, natural language processing (NLP), and dynamical systems. The results demonstrate that our proposed Taylor-Mixer, which replaces MLP and activation layers in the MLP-Mixer with Taylor layer, can achieve comparable accuracy on image classification, and similarly on sentiment analysis in NLP, while significantly reducing the number of model parameters. More importantly, our method can interpret some dynamical systems with Taylor polynomials. Meanwhile, the results demonstrate that our Taylor initialization can significantly improve classification accuracy compared to Xavier and Kaiming initialization.

****

Semi-supervised learning of partial differential operators and dynamical flows

Michael Rotman,Amit Dekel,Ran Ilan Ber,Lior Wolf,Yaron Oz

The evolution of dynamical systems is generically governed by nonlinear partial differential equations (PDEs), whose solution, in a simulation framework, requires vast amounts of computational resources. In this work, we present a novel method that combines a hyper-network solver with a Fourier Neural Operator architecture. Our method treats time and space separately and as a result, it successfully propagates initial conditions in continuous time steps by employing the general composition properties of the partial differential operators. Following previous works, supervision is provided at a specific time point. We test our method on various time evolution PDEs, including nonlinear fluid flows in one, two, or three spatial dimensions. The results show that the new method improves the learning accuracy at the time of the supervision point, and can interpolate the solutions to any intermediate time.

****

View Synthesis with Sculpted Neural Points

Yiming Zuo,Jia Deng

We address the task of view synthesis, generating novel views of a scene given a set of images as input. In many recent works such as NeRF (Mildenhall et al., 2020), the scene geometry is parameterized using neural implicit representations (i.e., MLPs). Implicit neural representations have achieved impressive visual quality but have drawbacks in computational efficiency. In this work, we propose a new approach that performs view synthesis using point clouds. It is the first point-based method that achieves better visual quality than NeRF while being 100× faster in rendering speed. Our approach builds on existing works on differentiable point-based rendering but introduces a novel technique we call “Sculpted Neural Points (SNP)”, which significantly improves the robustness to errors and holes in the reconstructed point cloud. We further propose to use view-dependent point features based on spherical harmonics to capture non-Lambertian surfaces, and new designs in the point-based rendering pipeline that further boost the performance. Finally, we show that our system supports fine-grained scene editing. Code is available at https://github.com/princeton-vl/SNP.

****

Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval

Zhenghao Liu,Chenyan Xiong,Yuanhuiyi Lv,Zhiyuan Liu,Ge Yu

This paper presents Universal Vision-Language Dense Retrieval (UniVL-DR), which builds a unified model for multi-modal retrieval. UniVL-DR encodes queries and multi-modality resources in an embedding space for searching candidates from different modalities. To learn a unified embedding space for multi-modal retrieval, UniVL-DR proposes two techniques: 1) Universal embedding optimization strategy, which contrastively optimizes the embedding space using the modality-balanced hard negatives; 2) Image verbalization method, which bridges the modality gap betw een images and texts in the raw data space. UniVL-DR achieves the state-of-the-art on the multi-modal open-domain question answering benchmark, WebQA, and outperforms all retrieval models on the two subtasks, text-text retrieval and text-image retrieval. It demonstrates that universal multi-modal search is feasible to replace the divide-and-conquer pipeline with a united model and also benefits single/cross modality tasks. All source codes of this work are available at https://github.com/OpenMatch/UniVL-DR.

****

DFlow: Learning to Synthesize Better Optical Flow Datasets via a Differentiable Pipeline

Kwon Byung-Ki,Nam Hyeon-Woo,Ji-Yun Kim,Tae-Hyun Oh

Comprehensive studies of synthetic optical flow datasets have attempted to reveal what properties lead to accuracy improvement in learning-based optical flow estimation. However, manually identifying and verifying the properties that contribute to accurate optical flow estimation require large-scale trial-and-error experiments with iteratively generating whole synthetic datasets and training on them, \ie, impractical. To address this challenge, we propose a differentiable optical flow data generation pipeline and a loss function to drive the pipeline, called DFlow. DFlow efficiently synthesizes a dataset effective for a target domain without the need for cumbersome try-and-errors. This favorable property is achieved by proposing an efficient dataset comparison method that uses neural networks to approximately encode each dataset and compares the proxy networks instead of explicitly comparing datasets in a pairwise way. Our experiments show the competitive performance of our DFlow against the prior arts in pre-training. Furthermore, compared to competing datasets, DFlow achieves the best fine-tuning performance on the Sintel public benchmark with RAFT.

****

One-Pixel Shortcut: On the Learning Preference of Deep Neural Networks

Shutong Wu,Sizhe Chen,Cihang Xie,Xiaolin Huang

Unlearnable examples (ULEs) aim to protect data from unauthorized usage for training DNNs. Existing work adds ℓ_∞-bounded perturbations to the original sample so that the trained model generalizes poorly. Such perturbations, however, are easy to eliminate by adversarial training and data augmentations. In this paper, we resolve this problem from a novel perspective by perturbing only one pixel in each image. Interestingly, such a small modification could effectively degrade model accuracy to almost an untrained counterpart. Moreover, our produced \emph{One-Pixel Shortcut (OPS)} could not be erased by adversarial training and strong augmentations. To generate OPS, we perturb in-class images at the same position to the same target value that could mostly and stably deviate from all the original images. Since such generation is only based on images, OPS needs significantly less computation cost than the previous methods using DNN generators. Based on OPS, we introduce an unlearnable dataset called CIFAR-10-S, which is indistinguishable from CIFAR-10 by humans but induces the trained model to extremely low accuracy. Even under adversarial training, a ResNet-18 trained on CIFAR-10-S has only 10.61% accuracy, compared to 83.02% by the existing error-minimizing method.

****

Make Memory Buffer Stronger in Continual Learning: A Continuous Neural Transformation Approach

Zhenyi Wang,Li Shen,Qiuling Suo,Tiehang Duan,Yanjun Zhu,Tongliang Liu,Mingchen Gao

Continual learning (CL) focuses on learning non-stationary data distribution without forgetting previous knowledge. However, the most widely used memory-replay approach often suffers from memory overfitting. To mitigate the memory overfitting, we propose a continuous and reversible memory transformation method so that the memory data is hard to overfit, thus improving generalization. The transformation is achieved by optimizing a bi-level optimization objective that jointly learns the CL model and memory transformer. Specifically, we propose a deterministic continuous memory transformer (DCMT) modeled by an ordinary differential equation, allowing for infinite memory transformation and generating diverse and ha rd memory data. Furthermore, we inject uncertainty into the transformation function and propose a stochastic continuous memory transformer (SCMT) modeled by a stochastic differential equation, which substantially enhances the diversity of the transformed memory buffer. The proposed neural transformation approaches have significant advantages over existing ones: (1) we can obtain infinite many transformed data, thus significantly increasing the memory buffer diversity; (2) the proposed continuous transformations are reversible, i.e., the original raw memory data could be restored from the transformed memory data without the need to make a replica of the memory data. Extensive experiments on both task-aware and task-free CL show significant improvement with our approach compared to strong baselines.

****

Sparse Random Networks for Communication-Efficient Federated Learning
Berivan Isik,Francesco Pase,Deniz Gunduz,Tsachy Weissman,Zorzi Michele

One main challenge in federated learning is the large communication cost of exchanging weight updates from clients to the server at each round. While prior work has made great progress in compressing the weight updates through gradient compression methods, we propose a radically different approach that does not update the weights at all. Instead, our method freezes the weights at their initial \emph{random} values and learns how to sparsify the random network for the best performance. To this end, the clients collaborate in training a \emph{stochastic} binary mask to find the optimal sparse random network within the original one. At the end of the training, the final model is a sparse network with random weights -- or a subnetwork inside the dense random network. We show improvements in accuracy, communication (less than 1 bit per parameter (bpp)), convergence speed, and final model size (less than 1 bpp) over relevant baselines on MNIST, EMNIST, CIFAR-10, and CIFAR-100 datasets, in the low bitrate regime.

****

On the Impact of Adversarially Robust Models on Algorithmic Recourse
Satyapriya Krishna,Chirag Agarwal,Himabindu Lakkaraju

The widespread deployment of machine learning models in various high-stakes settings has underscored the need for ensuring that individuals who are adversely impacted by model predictions are provided with a means for recourse. To this end, several algorithms have been proposed in recent literature to generate recourses. Recent research has also demonstrated that the recourses generated by these algorithms often correspond to adversarial examples. This key finding emphasizes the need for a deeper understanding of the impact of adversarially robust models (which are designed to guard against adversarial examples) on algorithmic recourse. In this work, we make one of the first attempts at studying the impact of adversarially robust models on algorithmic recourse. We theoretically and empirically analyze the cost (ease of implementation) and validity (probability of obtaining a positive model prediction) of the recourses output by state-of-the-art algorithms when the underlying models are adversarially robust. More specifically, we construct theoretical bounds on the differences between the cost and the validity of the recourses generated by various state-of-the-art algorithms when the underlying models are adversarially robust vs. non-robust. We also carry out extensive empirical analysis with multiple real-world datasets to not only validate our theoretical results, but also analyze the impact of varying degrees of model robustness on the cost and validity of the resulting recourses. Our theoretical and empirical analyses demonstrate that adversarially robust models significantly increase the cost and reduce the validity of the resulting recourses, thereby shedding light on the inherent trade-offs between achieving adversarial robustness in predictive models and providing easy-to-implement and reliable algorithmic recourse.

****

Breaking Beyond COCO Object Detection
ali borji

COCO dataset has become the de facto standard for training and evaluating object detectors. According to the recent benchmarks, however, performance on this dataset is still far from perfect, which raises the following questions, a) how

far can
we improve the accuracy on this dataset using deep learning, b) what is holding us back in making progress in object detection, and c) what are the limitations of the COCO dataset and how can they be mitigated. To answer these questions, first, we propose a systematic approach to determine the empirical upper bound in AP over COCOval2017, and show that this upper bound is significantly higher than the state-of-the-art mAP (78.2% vs. 58.8%). Second, we introduce two complementary datasets to COCO: i) COCO_OI, composed of images from COCO and OpenImages (from 80 classes in common) with 1,418,978 training bounding boxes over 380,111 images, and 41,893 validation bounding boxes over 18,299 images, and ii) ObjectNet_D containing objects in daily life situations (originally created for object recognition known as ObjectNet; 29 categories in common with COCO). We evaluate models on these datasets and pinpoint the annotation errors on the COCO validation set. Third, we characterize the sources of errors in modern object detectors using a recently proposed error analysis tool (TIDE) and find that models behave differently on these datasets compared to COCO. For instance, missing objects are more frequent in the new datasets. We also find that models lack out of distribution generalization. Code and data will be shared.

****

NormSoftmax: Normalize the Input of Softmax to Accelerate and Stabilize Training
Zixuan Jiang,Jiaqi Gu,David Z. Pan

Softmax is a basic function that normalizes a vector to a probability distribution and is widely used in machine learning, most notably in cross-entropy loss function and dot product attention operations. However, optimization of softmax-based models is sensitive to the input statistics change. We observe that the input of softmax changes significantly during the initial training stage, causing slow and unstable convergence when training the model from scratch. To remedy the optimization difficulty of softmax, we propose a simple yet effective substitution, named NormSoftmax, where the input vector is first normalized to unit variance and then fed to the standard softmax function. Similar to other existing normalization layers in machine learning models, NormSoftmax can stabilize and accelerate the training process, and also increase the robustness of the training procedure against hyperparameters. Experiments on Transformer-based models and convolutional neural networks validate that our proposed NormSoftmax is an effective plug-and-play module to stabilize and speed up the optimization of neural networks with cross-entropy loss or dot-product attention operations.

****

Differentially Private Dataset Condensation
Tianhang Zheng,Baochun Li

Recent work in ICML'22 builds a theoretical connection between dataset condensation (DC) and differential privacy (DP) and claims that DC can provide privacy protection for free. However, the connection is problematic because of two controversial assumptions. In this paper, we revisit the ICML'22 work and elucidate the issues in the two controversial assumptions. To correctly connect DC and DP, we propose two differentially private dataset condensation (DPDC) algorithms---LDPDC and NDPDC. Through extensive evaluations on multiple datasets, we demonstrate that LDPDC has comparable performance to recent DP generative methods despite its simplicity. NDPDC provides acceptable DP guarantees with a mild utility loss, compared to the state-of-the-art DC method. Additionally, NDPDC allows a flexible trade-off between the synthetic data utility and DP budget.

****

A General Framework For Proving The Equivariant Strong Lottery Ticket Hypothesis
Damien Ferbach,Christos Tsirigotis,Gauthier Gidel,Joey Bose

The Strong Lottery Ticket Hypothesis (SLTH) stipulates the existence of a subnetwork within a sufficiently overparameterized (dense) neural network that---when initialized randomly and without any training---achieves the accuracy of a fully trained target network. Recent works by Da Cunha et. al 2022, Burkholz 2022 dem onstrate that the SLTH can be extended to translation equivariant networks---i.e. CNNs---with the same level of overparametrization as needed for the SLTs in dense networks. However, modern neural networks are capable of incorporating more than just translation symmetry, and developing general equivariant architectures such as rotation and permutation has been a powerful design principle. In this paper, we generalize the SLTH to functions that preserve the action of the group G---i.e. G-equivariant network---and prove, with high probability, that one can approximate any G-equivariant network of fixed width and depth by pruning a randomly initialized overparametrized G-equivariant network to a G-equivariant subnetwork. We further prove that our prescribed overparametrization scheme is optimal and provide a lower bound on the number of effective parameters as a function of the error tolerance. We develop our theory for a large range of groups, including subgroups of the Euclidean $\text{E}(2)$ and Symmetric group $G \leq \mathcal{S}_n$---allowing us to find SLTs for MLPs, CNNs, $\text{E}(2)$-steerable CNNs, and permutation equivariant networks as specific instantiations of our unified framework. Empirically, we verify our theory by pruning overparametrized $\text{E}(2)$-steerable CNNs, k-order GNNs, and message passing GNNs to match the performance of trained target networks.

****

Robust Fair Clustering: A Novel Fairness Attack and Defense Framework

Anshuman Chhabra,Peizhao Li,Prasant Mohapatra,Hongfu Liu

Clustering algorithms are widely used in many societal resource allocation applications, such as loan approvals and candidate recruitment, among others, and hence, biased or unfair model outputs can adversely impact individuals that rely on these applications. To this end, many $\textit{fair}$ clustering approaches have been recently proposed to counteract this issue. Due to the potential for significant harm, it is essential to ensure that fair clustering algorithms provide consistently fair outputs even under adversarial influence. However, fair clustering algorithms have not been studied from an adversarial attack perspective. In contrast to previous research, we seek to bridge this gap and conduct a robustness analysis against fair clustering by proposing a novel $\textit{black-box fairness attack}$. Through comprehensive experiments, we find that state-of-the-art models are highly susceptible to our attack as it can reduce their fairness performance significantly. Finally, we propose Consensus Fair Clustering (CFC), the first $\textit{robust fair clustering}$ approach that transforms consensus clustering into a fair graph partitioning problem, and iteratively learns to generate fair cluster outputs. Experimentally, we observe that CFC is highly robust to the proposed attack and is thus a truly robust fair clustering alternative.

****

Learning to Jointly Share and Prune Weights for Grounding Based Vision and Language Models

Shangqian Gao,Burak Uzkent,Yilin Shen,Heng Huang,Hongxia Jin

Transformers have seen growing interest in processing different modalities, including language and image data. As a result, we can process vision and language data using transformers that are architecturally similar. Leveraging this feature of transformers, we propose weight sharing across two transformer backbones and within the same transformer backbone and pruning across two backbones in a unified framework. More specifically, we investigate weight sharing and pruning for two components of the transformers: (1) Multi-Head Attention (MSA) and (2) Feed-Forward Network (FFN) layers. To jointly perform weight sharing and pruning, we propose to use a regularization term to align model weights and the desired structure during the multimodal pre-training step. The structure vectors of sharing and pruning are generated by using a hypernetwork, which can capture complex interactions between pruning and sharing across layers and modalities. We train the hypernetwork and model weights iteratively so that the learned structure evolves along with model weights. After minimizing the proposed objective in the pre-training step, we perform weight sharing and pruning and fine-tune the compressed model on downstream tasks. Finally, we perform experiments on vision and language tasks, including Referring Expression Comprehension (REC), Visual Question Answering (VQA), and Object Detection using the state-of-the-art grounding based models: MDETR and GLIP. Our experiments show that we can compress these models by $35-40\%$ by sharing and pruning MSA and FFN weights without almost any loss in accuracy.

****

Spatial Attention Kinetic Networks with E(n)-Equivariance

Yuanqing Wang,John Chodera

Neural networks that are equivariant to rotations, translations, reflections, and permutations on n-dimensional geometric space have shown promise in physical modeling for tasks such as accurately but inexpensively modeling complex potential energy surfaces to guiding the sampling of complex dynamical systems or forecasting their time evolution.

Current state-of-the-art methods employ spherical harmonics to encode higher-order interactions among particles, which are computationally expensive.

In this paper, we propose a simple alternative functional form that uses neurally parametrized linear combinations of edge vectors to achieve equivariance while still universally approximating node environments.

Incorporating this insight, we design \emph{spatial attention kinetic networks} with E(n)-equivariance, or SAKE, which are competitive in many-body system modeling tasks while being significantly faster.

****

Light-weight probing of unsupervised representations for Reinforcement Learning

Wancong Zhang,Anthony GX-Chen,Vlad Sobal,Yann LeCun,Nicolas Carion

Unsupervised visual representation learning offers the opportunity to leverage large corpora of unlabeled trajectories to form useful visual representations, which can benefit the training of reinforcement learning (RL) algorithms. However, evaluating the fitness of such representations requires training RL algorithms which is both computationally intensive and has high variance outcomes. To alleviate this issue, we design an evaluation protocol for unsupervised RL representations with lower variance and up to 600x lower computational cost. Inspired by the vision community, we propose two linear probing tasks: predicting the reward observed in a given state, and predicting the action of an expert in a given state. These two tasks are generally applicable to many RL domains, and we show through rigorous experimentation that they correlate strongly with the actual downstream control performance on the Atari100k Benchmark. This provides a better method for exploring the space of pretraining algorithms without the need of running RL evaluations for every setting. Leveraging this framework, we further improve existing self-supervised learning (SSL) recipes for RL, highlighting the importance of the forward model, the size of the visual backbone, and the precise formulation of the unsupervised objective.

****

Understanding Rare Spurious Correlations in Neural Networks

Yao-Yuan Yang,Chi-Ning Chou,Kamalika Chaudhuri

Neural networks are known to use spurious correlations such as background information for classification. While prior work has looked at spurious correlations that are widespread in the training data, in this work, we investigate how sensitive neural networks are to $rare$ spurious correlations, which may be harder to detect and correct, and may lead to privacy leaks. We introduce spurious patterns correlated with a fixed class to a few training examples and find that it takes only a handful of such examples for the network to learn the correlation. Furthermore, these rare spurious correlations also impact accuracy and privacy. We empirically and theoretically analyze different factors involved in rare spurious correlations and propose mitigation methods accordingly. Specifically, we observe that ℓ_2 regularization and adding Gaussian noise to inputs can reduce the undesirable effects.

****

Graph Domain Adaptation via Theory-Grounded Spectral Regularization

Yuning You,Tianlong Chen,Zhangyang Wang,Yang Shen

Transfer learning on graphs drawn from varied distributions (domains) is in great demand across many applications. Emerging methods attempt to learn domain-invariant representations using graph neural networks (GNNs), yet the empirical perf ormances vary and the theoretical foundation is limited. This paper aims at designing theory-grounded algorithms for graph domain adaptation (GDA). (i) As the first attempt, we derive a model-based GDA bound closely related to two GNN spectral properties: spectral smoothness (SS) and maximum frequency response (MFR). This is achieved by cross-pollinating between the OT-based (optimal transport) DA and graph filter theories. (ii) Inspired by the theoretical results, we propose algorithms regularizing spectral properties of SS and MFR to improve GNN transferability. We further extend the GDA theory into the more challenging scenario of conditional shift, where spectral regularization still applies. (iii) More importantly, our analyses of the theory reveal which regularization would improve performance of what transfer learning scenario, (iv) with numerical agreement with extensive real-world experiments: SS and MFR regularizations bring more benefits to the scenarios of node transfer and link transfer, respectively. In a nutshell, our study paves the way toward explicitly constructing and training GNNs that can capture more transferable representations across graph domains. Codes are released at https://github.com/Shen-Lab/GDA-SpecReg.

****

Effective dimension of machine learning models

Amira Abbas,David Sutter,Alessio Figalli,Stefan Woerner

Making statements about the performance of trained models on tasks involving new data is one of the primary goals of machine learning, i.e., to understand the generalization power of a model. Various capacity measures try to capture this ability, but usually fall short in explaining important characteristics of models that we observe in practice. In this study, we propose the local effective dimension as a capacity measure which seems to correlate well with generalization error on standard data sets. Importantly, we prove that the local effective dimension bounds the generalization error and discuss the aptness of this capacity measure for machine learning models.

****

CLARE: Conservative Model-Based Reward Learning for Offline Inverse Reinforcement Learning

Sheng Yue,Guanbo Wang,Wei Shao,Zhaofeng Zhang,Sen Lin,Ju Ren,Junshan Zhang

This work aims to tackle a major challenge in offline Inverse Reinforcement Learning (IRL), namely the reward extrapolation error, where the learned reward function may fail to explain the task correctly and misguide the agent in unseen environments due to the intrinsic covariate shift. Leveraging both expert data and lower-quality diverse data, we devise a principled algorithm (namely CLARE) that solves offline IRL efficiently via integrating "conservatism" into a learned reward function and utilizing an estimated dynamics model. Our theoretical analysis provides an upper bound on the return gap between the learned policy and the expert policy, based on which we characterize the impact of covariate shift by examining subtle two-tier tradeoffs between the exploitation (on both expert and diverse data) and exploration (on the estimated dynamics model). We show that CLARE can provably alleviate the reward extrapolation error by striking the right exploitation-exploration balance therein. Extensive experiments corroborate the significant performance gains of CLARE over existing state-of-the-art algorithms on MuJoCo continuous control tasks (especially with a small offline dataset), and the learned reward is highly instructive for further learning.

****

Data-Free One-Shot Federated Learning Under Very High Statistical Heterogeneity

Clare Elizabeth Heinbaugh,Emilio Luz-Ricca,Huajie Shao

Federated learning (FL) is an emerging distributed learning framework that collaboratively trains a shared model without transferring the local clients' data to a centralized server. Motivated by concerns stemming from extended communication and potential attacks, one-shot FL limits communication to a single round while attempting to retain performance. However, one-shot FL methods often degrade under high statistical heterogeneity, fail to promote pipeline security, or require an auxiliary public dataset. To address these limitations, we propose two novel data-free one-shot FL methods: FedCVAE-Ens and its extension FedCVAE-KD. Both approaches reframe the local learning task using a conditional variational auto

encoder (CVAE) to address high statistical heterogeneity. Furthermore, FedCVAE-KD leverages knowledge distillation to compress the ensemble of client decoders into a single decoder. We propose a method that shifts the center of the CVAE prior distribution and experimentally demonstrate that this promotes security, and show how either method can incorporate heterogeneous local models. We confirm the efficacy of the proposed methods over baselines under high statistical heterogeneity using multiple benchmark datasets. In particular, at the highest levels of statistical heterogeneity, both FedCVAE-Ens and FedCVAE-KD typically more than double the accuracy of the baselines.
****

Personalized Subgraph Federated Learning
Jinheon Baek,Wonyong Jeong,Jiongdao Jin,Jaehong Yoon,Sung Ju Hwang
In real-world scenarios, subgraphs of a larger global graph may be distributed across multiple devices or institutions, and only locally accessible due to privacy restrictions, although there may be links between them. Recently proposed subgraph Federated Learning (FL) methods deal with those missing links across private local subgraphs while distributively training Graph Neural Networks (GNNs) on them. However, they have overlooked the inevitable heterogeneity among subgraphs, caused by subgraphs comprising different communities of a global graph, therefore, consequently collapsing the incompatible knowledge from local GNN models trained on heterogeneous graph distributions. To overcome such a limitation, we introduce a new subgraph FL problem, personalized subgraph FL, which focuses on the joint improvement of the interrelated local GNN models rather than learning a single global GNN model, and propose a novel framework, FEDerated Personalized sUBgraph learning (FED-PUB), to tackle it. A crucial challenge in personalized subgraph FL is that the server does not know which subgraph each client has. FED-PUB thus utilizes functional embeddings of the local GNNs using random graphs as inputs to compute similarities between them, and use them to perform weighted averaging for server-side aggregation. Further, it learns a personalized sparse mask at each client to select and update only the subgraph-relevant subset of the aggregated parameters. We validate FED-PUB for its subgraph FL performance on six datasets, considering both non-overlapping and overlapping subgraphs, on which ours largely outperforms relevant baselines.
****

Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization
Elan Rosenfeld,Pradeep Kumar Ravikumar,Andrej Risteski
A common explanation for the failure of deep networks to generalize out-of-distribution is that they fail to recover the "correct" features. We challenge this notion with a simple experiment which suggests that ERM already learns sufficient features and that the current bottleneck is not feature learning, but robust regression. We therefore argue that devising simpler methods for learning predictors on existing features is a promising direction for future research. Towards this end, we introduce Domain-Adjusted Regression (DARE), a convex objective for learning a linear predictor that is provably robust under a new model of distribution shift. Rather than learning one function, DARE performs a domain-specific adjustment to unify the domains in a canonical latent space and learns to predict in this space. Under a natural model, we prove that the DARE solution is the minimax-optimal predictor for a constrained set of test distributions. Further, we provide the first finite-environment convergence guarantee to the minimax risk, improving over existing analyses which only yield minimax predictors after an environment threshold. Evaluated on finetuned features, we find that DARE compares favorably to prior methods, consistently achieving equal or better performance.
****

Initial Value Problem Enhanced Sampling for Closed-Loop Optimal Control Design with Deep Neural Networks
Xuanxi Zhang,Jihao Long,Wei Hu,Weinan E,Jiequn Han
Closed-loop optimal control design for high-dimensional nonlinear systems has been a long-standing challenge. Traditional methods, such as solving the associate

d Hamilton-Jacobi-Bellman equation, suffer from the curse of dimensionality. Recent literature proposed a new promising approach based on supervised learning, by leveraging powerful open-loop optimal control solvers to generate training data and neural networks as efficient high-dimensional function approximators to fit the closed-loop optimal control. This approach successfully handles certain high-dimensional optimal control problems but still performs poorly on more challenging problems. One of the crucial reasons for the failure is the so-called distribution mismatch phenomenon brought by the controlled dynamics. In this paper, we investigate this phenomenon and propose the initial value problem enhanced sampling method to mitigate this problem. We theoretically prove that this sampling strategy improves over the vanilla strategy on the classical linear-quadratic regulator by a factor proportional to the total time duration. We further numerically demonstrate that the proposed sampling strategy significantly improves the performance on tested control problems, including the optimal landing problem of a quadrotor and the optimal reaching problem of a 7 DoF manipulator.

****

C3PO: Learning to Achieve Arbitrary Goals via Massively Entropic Pretraining
Alexis D. Jacq,Manu Orsini,Gabriel Dulac-Arnold,Olivier Pietquin,Matthieu Geist,Olivier Bachem

Given a particular embodiment, we propose a novel method (C3PO) that learns policies able to achieve any arbitrary position and pose. Such a policy would allow for easier control, and would be re-useable as a key building block for downstream tasks. The method is two-fold: First, we introduce a novel exploration algorithm that optimizes for uniform coverage, is able to discover a set of achievable states, and investigates its abilities in attaining both high coverage, and hard-to-discover states; Second, we leverage this set of achievable states as training data for a universal goal-achievement policy, a goal-based SAC variant. We demonstrate the trained policy's performance in achieving a large number of novel states. Finally, we showcase the influence of massive unsupervised training of a goal-achievement policy with state-of-the-art pose-based control of the Hopper, Walker, Halfcheetah, Humanoid and Ant embodiments.

****

ProtoVAE: Using Prototypical Networks for Unsupervised Disentanglement
Vaishnavi S Patil,Matthew Evanusa,Joseph JaJa

Generative modeling and self-supervised learning have in recent years made great strides towards learning from data in a completely \emph{unsupervised} way. There is still, however, an open area of investigation into guiding the neural network to learn useful or good representations. The problem of unsupervised \textit{Disentanglement} is of particular importance as it offers to learn interpretable representations, with disjoint subsets of the representation encoding different, meaningful factors of variation. Recent work has theoretically grounded the factors of variation, via the lens of group theory, as disentangled actions of the symmetry subgroups which transform only the correspond subspaces of the disentangled representation. We use this mathematical formalism instead to impose constraints on the representations learned by a unsupervised generative neural network, such that transformations of the representation correspond to the actions of a unique symmetry subgroup. To this end, we introduce a novel model, ProtoVAE, that leverages a deep metric learning Prototypical network trained via self-supervision to constrain the latent space of a Variational Autoencoder to decompose into independent subspaces. Further, we actively change or \textit{intervene} in the latent space during training to enforce each dimension of the representation to uniquely and consistently transform the data corresponding to some symmetry subgroup. We demonstrate and evaluate our proposed model on the benchmark DSprites and 3DShapes datasets and compare with other state of the art disentanglement methods via qualitative traversals in the latent space, as well as quantitative disentanglement metrics. We further qualitatively demonstrate the effectiveness of our model on the real-world datasets CelebA which consistently encodes the different factors.

****

Neural Diffusion Processes

Vincent Dutordoir,Alan Saul,Zoubin Ghahramani,Fergus Simpson

Gaussian processes provide an elegant framework for specifying prior and posterior distributions over functions. They are, however, also computationally expensive, and limited by the expressivity of their covariance function. We propose Neural Diffusion Processes (NDPs), a novel approach based upon diffusion models, that learns to sample from distributions over functions. Using a novel attention block we are able to incorporate properties of stochastic processes, such as exchangeability, directly into the NDP's architecture. We empirically show that NDPs are able to capture functional distributions that are close to the true Bayesian posterior. This enables a variety of downstream tasks, including hyperparameter marginalisation, non-Gaussian posteriors and global optimisation.

****

Global Context Vision Transformers

Ali Hatamizadeh,Hongxu Yin,Jan Kautz,Pavlo Molchanov

We propose global context vision transformer (GC ViT), a novel architecture that enhances parameter and compute utilization for computer vision tasks. The core of the novel model are global context self-attention modules, joint with standard local self-attention, to effectively yet efficiently model both long and short-range spatial interactions, as an alternative to complex operations such as an attention masks or local windows shifting. While the local self-attention modules are responsible for modeling short-range information, the global query tokens are shared across all global self-attention modules to interact with local key and values. In addition, we address the lack of inductive bias in ViTs and improve the modeling of inter-channel dependencies by proposing a novel downsampler which leverages a parameter-efficient fused inverted residual block. The proposed GC ViT achieves new state-of-the-art performance across image classification, object detection and semantic segmentation tasks. On ImageNet-1K dataset for classification, the tiny, small and base variants of GC ViT with 28M, 51M and 90M parameters achieve 83.4%, 83.9% and 84.4% Top-1 accuracy, respectively, surpassing comparably-sized prior art such as CNN-based ConvNeXt and ViT-based Swin Transformer. Pre-trained GC ViT backbones in downstream tasks of object detection, instance segmentation, and semantic segmentation on MS COCO and ADE20K datasets outperform prior work consistently, sometimes by large margins.

****

Adversarial Learned Fair Representations using Dampening and Stacking

Max Knobbout

As more decisions in our daily life become automated, the need to have machine learning algorithms that make fair decisions increases. In fair representation learning we are tasked with finding a suitable representation of the data in which a sensitive variable is censored. Recent work aims to learn fair representations through adversarial learning. This paper builds upon this work by introducing a novel algorithm which uses dampening and stacking to learn adversarial fair representations. Results show that that our algorithm improves upon earlier work in both censoring and reconstruction.

****

Watch What You Pretrain For: Targeted, Transferable Adversarial Examples on Self-Supervised Speech Recognition models

Raphael Olivier,Hadi Abdullah,Bhiksha Raj

A targeted adversarial attack produces audio samples that can force an Automatic Speech Recognition (ASR) system to output attacker-chosen text. To exploit ASR models in real-world, black-box settings, an adversary can leverage the transferability property, i.e. that an adversarial sample produced for a proxy ASR can also fool a different remote ASR. However recent work has shown that transferability against large ASR models is very difficult. In this work, we show that modern ASR architectures, specifically ones based on Self-Supervised Learning, are in fact vulnerable to transferability. We successfully demonstrate this phenomenon by evaluating state-of-the-art self-supervised ASR models like Wav2Vec2, HuBERT, Data2Vec and WavLM. We show that with low-level additive noise achieving a 30dB Signal-Noise Ratio, we can achieve target transferability with up to 80\% accuracy. Next, we 1) use an ablation study to show that Self-Supervised learning i s the main cause of that phenomenon, and 2) we provide an explanation for this phenomenon. Through this we show that modern ASR architectures are uniquely vulnerable to adversarial security threats.

****

Imposing conservation properties in deep dynamics modeling via contrastive learning

Wang Zhang,Subhro Das,Tsui-Wei Weng,Alexandre Megretski,Luca Daniel,Lam M. Nguyen

Deep neural networks (DNN) has shown great capacity of modeling a dynamical system, but these DNN-based dynamical models usually do not obey conservation laws. To impose the learned DNN dynamical models with key physical properties such as conservation laws, this paper proposes a two-step approach to endow the invariant priors into the simulations. We first establish a contrastive learning framework to capture the system invariants along the trajectory observations. During the dynamics modeling, we design a projection layer of DNNs to preserve the system invariance. Through experiments, we show our method consistently outperforms the baseline in both coordinate error and conservation metrics and can be further extended to complex and large dynamics by leveraging autoencoder. Notably, a byproduct of our framework is the automated conservation law discovery for dynamical systems with single conservation property.

****

GReTo: Remedying dynamic graph topology-task discordance via target homophily

Zhengyang Zhou,qihe huang,Gengyu Lin,Kuo Yang,LEI BAI,Yang Wang

Dynamic graphs are ubiquitous across disciplines where observations usually change over time. Regressions on dynamic graphs often contribute to diverse critical tasks, such as climate early-warning and traffic controlling. Existing homophily Graph Neural Networks (GNNs) adopt physical connections or feature similarity as adjacent matrix to perform node-level aggregations. However, on dynamic graphs with diverse node-wise relations, exploiting a pre-defined fixed topology for message passing inevitably leads to the aggregations of target-deviated neighbors. We designate such phenomenon as the topology-task discordance, which naturally challenges the homophily assumption. In this work, we revisit node-wise relationships and explore novel homophily measurements on dynamic graphs with both signs and distances, capturing multiple node-level spatial relations and temporal evolutions. We discover that advancing homophily aggregations to signed target-oriented message passing can effectively resolve the discordance and promote aggregation capacity. Therefore, a GReTo is proposed, which performs signed message passing in immediate neighborhood, and exploits both local environments and target awareness to realize high-order message propagation. Empirically, our solution achieves significant improvements against best baselines, notably improving 24.79% on KnowAir and 3.60% on Metr-LA.

****

Towards predicting dynamic stability of power grids with Graph Neural Networks

Christian Nauck,Michael Lindner,Konstantin Schürholt,Frank Hellmann

To mitigate climate change, the share of renewable energies in power production needs to be increased. Renewables introduce new challenges to power grids regarding the dynamic stability due to decentralization, reduced inertia and volatility in production. However, dynamic stability simulations are intractable and exceedingly expensive for large grids. Graph Neural Networks (GNNs) are a promising method to reduce the computational effort of analyzing dynamic stability of power grids. We provide new datasets of dynamic stability of synthetic power grids and find that GNNs are surprisingly effective at predicting highly non-linear targets from topological information only. We show that large GNNs outperform GNNs from previous work as well as as handcrafted graph features and semi-analytic approximations. Further, we demonstrate GNNs can accurately identify \emph{trouble maker}-nodes in the power grids. Lastly, we show that GNNs trained on small grids can perform accurately on a large synthetic Texan power grid model, which illustrates the potential of our approach.

****

Deep Reinforcement Learning for Cost-Effective Medical Diagnosis

Zheng Yu,Yikuan Li,Joseph Chahn Kim,Kaixuan Huang,Yuan Luo,Mengdi Wang

Dynamic diagnosis is desirable when medical tests are costly or time-consuming. In this work, we use reinforcement learning (RL) to find a dynamic policy that selects lab test panels sequentially based on previous observations, ensuring accurate testing at a low cost. Clinical diagnostic data are often highly imbalanced; therefore, we aim to maximize the F1 score instead of the error rate. However, optimizing the non-concave F_1 score is not a classic RL problem, thus invalidating standard RL methods. To remedy this issue, we develop a reward shaping approach, leveraging properties of the F_1 score and duality of policy optimization, to provably find the set of all Pareto-optimal policies for budget-constrained F_1 score maximization. To handle the combinatorially complex state space, we propose a Semi-Model-based Deep Diagnosis Policy Optimization (SM-DDPO) framework that is compatible with end-to-end training and online learning. SM-DDPO is tested on diverse clinical tasks: ferritin abnormality detection, sepsis mortality prediction, and acute kidney injury diagnosis. Experiments with real-world data validate that SM-DDPO trains efficiently and identify all Pareto-front solutions. Across all tasks, SM-DDPO is able to achieve state-of-the-art diagnosis accuracy (in some cases higher than conventional methods) with up to 85% reduction in testing cost. Core codes are available at https://github.com/Zheng321/Deep-Reinforcement-Learning-for-Cost-Effective-Medical-Diagnosis.

****

Deterministic training of generative autoencoders using invertible layers

Gianluigi Silvestri,Daan Roos,Luca Ambrogioni

In this work, we provide a deterministic alternative to the stochastic variational training of generative autoencoders. We refer to these new generative autoencoders as AutoEncoders within Flows (AEF), since the encoder and decoder are defined as affine layers of an overall invertible architecture. This results in a deterministic encoding of the data, as opposed to the stochastic encoding of VAEs. The paper introduces two related families of AEFs. The first family relies on a partition of the ambient space and is trained by exact maximum-likelihood. The second family exploits a deterministic expansion of the ambient space and is trained by maximizing the log-probability in this extended space. This latter case leaves complete freedom in the choice of encoder, decoder and prior architectures, making it a drop-in replacement for the training of existing VAEs and VAE-style models. We show that these AEFs can have strikingly higher performance than architecturally identical VAEs in terms of log-likelihood and sample quality, especially for low dimensional latent spaces. Importantly, we show that AEF samples are substantially sharper than VAE samples.

****

ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging

Alessandro Fontanella,Antreas Antoniou,Wenwen Li,Joanna Wardlaw,Grant Mair,Emanuele Trucco,Amos Storkey

In some medical imaging tasks and other settings where only small parts of the image are informative for the classification task, traditional CNNs can sometimes struggle to generalise. Manually annotated Regions of Interest (ROI) are sometimes used to isolate the most informative parts of the image. However, these are expensive to collect and may vary significantly across annotators. To overcome these issues, we propose a method to generate saliency maps, obtained from adversarially generated counterfactual images. With this method, we are able to isolate the area of interest in brain and lung CT scans without using any manual annotations. Our saliency maps, in the task of localising the lesion location out of 6 possible regions, obtain a score of 65.05% on brain CT scans, improving the score of 61.29% obtained with the best competing method. We also employ the saliency maps in a framework that refines a classifier pipeline. In particular, the saliency maps are used to obtain soft spatial attention masks that modulate the image features at different scales. We refer to our method as \emph{Adversarial Counterfactual Attention} (ACAT). ACAT increases the baseline classification accuracy of lesions in brain CT scans from 71.39% to 72.55% and of COVID-19 related findings in lung CT scans from 67.71% to 70.84% and excee

ds the performance of competing methods.
****

Abstract Visual Reasoning by Self-supervised Contrastive Learning

Weiwen Lu,Aihua Yin,Sidong Wang,Hongzhi You,Ru-Yuan Zhang,Dahui Wang,Zonglei Zhen,Xiaohong Wan

Neuro-symbolic models of artificial intelligence (AI) have been recently developed to perform tasks involving abstract visual reasoning that is a hallmark of human intelligence but remains challenging for deep neural network methods. However, most of the current neuro-symbolic models also rely on supervised learning and auxiliary annotations, different from human cognitive processes that are much dependent on the general cognitive abilities of entity and rule recognitions, rather than learning how to solve the specific tasks from examples. In this work, we propose a neuro-symbolic model by self-supervised contrastive learning (NS-SSCL) with unique and invariant representations of entities and rules in the perception and reasoning modules, respectively, to solve Raven’s Progressive Matrices (RPMs) and its variant, a typical type of visual reasoning task used to test human intelligence. The perception module parses each object into invariant representations of attributes. The reasoning module grounds the representations of object attributes to form the latent rule representations also through SSCL. Further, the relationships between the neural representations of object attributes and symbols used for rule reasoning are coherently mapped. Finally, the scene generation engine aggregates all attribute and rule representation distributions to produce a probabilistic representation of the target. NS-SSCL obtains state-of-the-art performance in unsupervised models to solve the RAVEN and V-PROM benchmarks, even better than most of the supervised models. The success of the proposed model suggests that construction of general cognitive abilities like humans may render the AI algorithms to solve complex tasks involving higher-level cognition such as abstract reasoning in a human-like manner.
****

POPGym: Benchmarking Partially Observable Reinforcement Learning

Steven Morad,Ryan Kortvelesy,Matteo Bettini,Stephan Liwicki,Amanda Prorok

Real world applications of Reinforcement Learning (RL) are often partially observable, thus requiring memory. Despite this, partial observability is still largely ignored by contemporary RL benchmarks and libraries. We introduce Partially Observable Process Gym (POPGym), a two-part library containing (1) a diverse collection of 15 partially observable environments, each with multiple difficulties and (2) implementations of 13 memory model baselines -- the most in a single RL library. Existing partially observable benchmarks tend to fixate on 3D visual navigation, which is computationally expensive and only one type of POMDP. In contrast, POPGym environments are diverse, produce smaller observations, use less memory, and often converge within two hours of training on a consumer-grade GPU. We implement our high-level memory API and memory baselines on top of the popular RLlib framework, providing plug-and-play compatibility with various training algorithms, exploration strategies, and distributed training paradigms. Using POPGym, we execute the largest comparison across RL memory models to date. POPGym is available at https://github.com/proroklab/popgym.
****

HierBatching: Locality-Aware Out-of-Core Training of Graph Neural Networks

Tianhao Huang,Xuhao Chen,Muhua Xu,Arvind Arvind,Jie Chen

As graph neural networks (GNNs) become increasingly more popular for analyzing data organized as massive graphs, how these models can be efficiently trained under economic computing resources becomes a critical subject that influences the widespread adoption of GNNs in practice. We consider the use of a single commodity machine restrained by limited memory but otherwise is attached with ample external storage. In such an under-explored scenario, not only the feature data often exceeds the memory capacity, but also the graph structure may not fit in memory as well. Then, with data stored on disk, gathering features and constructing neighborhood subgraphs in a usual mini-batch training incur inefficient random access and expensive data movement.

To overcome this bottleneck, we propose a locality-aware training scheme, coined HierBatching, to significantly increase sequential disk access, while maintaining the random nature of stochastic training and its quality. HierBatching exploits the memory hierarchy of a modern GPU machine and constructs batches in an analogously hierarchical manner. Therein, graph nodes are organized in many partitions, each of which is laid out contiguously in disk for maximal spatial locality; while the main memory stores random partitions and is treated as the cache of the disk. Its content is reused multiple times for improving temporal locality. We conduct comprehensive experiments, including locality ablation, to demonstrate that HierBatching is economic, fast, and accurate.
****

Everybody Needs Good Neighbours: An Unsupervised Locality-based Method for Bias Mitigation
Xudong Han,Timothy Baldwin,Trevor Cohn
Learning models from human behavioural data often leads to outputs that are biased with respect to user demographics, such as gender or race. This effect can be controlled by explicit mitigation methods, but this typically presupposes access to demographically-labelled training data. Such data is often not available, motivating the need for unsupervised debiasing methods. To this end, we propose a new meta-algorithm for debiasing representation learning models, which combines the notions of data locality and accuracy of model fit, such that a supervised debiasing method can optimise fairness between neighbourhoods of poorly vs. well modelled instances as identified by our method. Results over five datasets, spanning natural language processing and structured data classification tasks, show that our technique recovers proxy labels that correlate with unknown demographic data, and that our method outperforms all unsupervised baselines, while also achieving competitive performance with state-of-the-art supervised methods which are given access to demographic labels.
****

Continual Learning In Low-coherence Subspace: A Strategy To Mitigate Learning Capacity Degradation
Zihao Xu,Zhengyu Li,Yufei Shi,Xihao Wang,Yingjie Liu,Kexin Ke,Xian Wei
Methods using gradient orthogonal projection, an efficient strategy in continual learning, have achieved promising success in mitigating catastrophic forgetting. However, these methods often suffer from the learning capacity degradation problem following the increasing number of tasks. To address this problem, we propose to learn new tasks in low-coherence subspaces rather than orthogonal subspaces. Specifically, we construct a unified cost function involving regular DNN parameters and gradient projections on the Oblique manifold. We finally develop a gradient descent algorithm on a smooth manifold to jointly minimize the cost function and minimize both the inter-task and the intra-task coherence. Numerical experimental results show that the proposed method has prominent advantages in maintaining the learning capacity when tasks are increased, especially on a large number of tasks, compared with baselines.
****

Particle-based Variational Inference with Preconditioned Functional Gradient Flow
Hanze Dong,Xi Wang,LIN Yong,Tong Zhang
Particle-based variational inference (VI) minimizes the KL divergence between model samples and the target posterior with gradient flow estimates. With the popularity of Stein variational gradient descent (SVGD), the focus of particle-based VI algorithms has been on the properties of functions in Reproducing Kernel Hilbert Space (RKHS) to approximate the gradient flow. However, the requirement of RKHS restricts the function class and algorithmic flexibility. This paper offers a general solution to this problem by introducing a functional regularization term that encompasses the RKHS norm as a special case. This allows us to propose a new particle-based VI algorithm called preconditioned functional gradient flow (PFG). Compared to SVGD, PFG has several advantages. It has a larger function class, improved scalability in large particle-size scenarios, better adaptation to ill-conditioned distributions, and provable continuous-time convergence in K

L divergence. Additionally, non-linear function classes such as neural networks can be incorporated to estimate the gradient flow. Our theory and experiments demonstrate the effectiveness of the proposed framework.

****

An Efficient Mean-field Approach to High-Order Markov Logic

Weidi Xu,Jianshan He,Jingwei Wang,Hongting Zhou,Xiaopei Wan,Taifeng Wang,Ruopeng Li,Wei Chu

Markov logic networks (MLNs) are powerful models for symbolic reasoning, which combine probabilistic modeling with relational logic. Inference algorithms for MLNs often perform at the level of propositional logic or require building a first-order probabilistic graph, and the computational efficiency remains a challenge. The mean-field algorithm generalizes message passing for approximate inference in many intractable probabilistic graphical models, but in MLNs it still suffers from the high-order dependencies among the massive groundings, resulting in time complexity exponential in both the length and the arity of logic rules. We propose a novel method, LogicMP, to simplify the logic message passing especially. In most practical cases, it can reduce the complexity significantly to polynomial for the formulae in conjunctive normal form (CNF). We exploit the property of CNF logic rules to sidestep the expectation computation of high-order dependency, and then formulate the logic message passing by Einstein summation to facilitate parallel computation, which can be optimized by sequentially contracting the rule arguments. With LogicMP, we achieve evident improvements on several reasoning benchmark datasets in both performance and efficiency over competitor methods. Specifically, the AUC-PR of the UW-CSE and Cora datasets is improved by more than 11\% absolutely and the speed is about ten times faster.

****

A theory of representation learning in neural networks gives a deep generalisation of kernel methods

Adam X. Yang,Maxime Robeyns,Edward Milsom,Nandi Schoots,Laurence Aitchison

The successes of modern deep neural networks (DNNs) are founded on their ability to transform inputs across multiple layers to build good high-level representations. It is therefore critical to understand this process of representation learning. However, standard theoretical approaches (formally NNGPs) involving infinite width limits eliminate representation learning. We therefore develop a new infinite width limit, the representation learning limit, that exhibits representation learning mirroring that in finite-width networks, yet at the same time, remains extremely theoretically tractable. In particular, we derive an elegant objective that describes how each network layer learns representations that interpolate between input and output, and we confirm that the modes of the objective match the behaviour of finite but wide networks. Moreover, we use this limit and objective to develop a flexible, deep generalisation of kernel methods, that we call deep kernel machines (DKMs). We show that DKMs can be scaled to large datasets using inducing point methods from the Gaussian process literature, and we show that DKMs exhibit superior performance to other kernel-based approaches.

****

A spatiotemporal graph neural network with multi granularity for air quality prediction

Haibin Liao,Yuan Li,Mou Wu,Hongsheng Chen

Air quality prediction is a complex system engineering. How to fully consider the impact of meteorological, spatial and temporal factors on air quality is the core problem. To address this central conundrum, in an elaborate encoder-decoder architecture, we propose a new air quality prediction method based on multi-granularity spatiotemporal graph network. At the encoder, firstly, we use multi granularity graph and the well-known HYSPLIT model to build spatial relationship and dynamic edge relationship between nodes, respectively, while meteorological, temporal and topographic characteristics are used to build node features and LSTM (Long Short Term Memory) is used to learn the time-series relationship of pollutant concentration. At the decoder, secondly, we use the attention mechanism LSTM for decoding and forecasting of pollutant concentration. The proposed model is capable of tracking different influences on prediction resulting from the changes of air quality. On a project-based dataset, we validate the effectiveness of the proposed model and examine its abilities of capturing both fine-grained and long-term influences in pollutant process. We also compare the proposed model with the state-of-the-art air quality forecasting methods on the dataset of Yangtze River Delta city group, the experimental results show the appealing performance of our model over competitive baselines.

****

Highway Reinforcement Learning

Yuhui Wang,Haozhe Liu,Miroslav Strupl,Francesco Faccio,Qingyuan Wu,Xiaoyang Tan,Jürgen Schmidhuber

Traditional Dynamic Programming (DP) approaches suffer from slow backward credit-assignment (CA): only a one-step search is performed at each update. A popular solution for multi-step CA is to use multi-step Bellman operators. Unfortunately, in the control settings, existing methods typically suffer from the large variance of multi-step off-policy corrections or are biased, preventing convergence. To overcome these problems, we introduce a novel multi-step Bellman optimality equation with adaptive lookahead steps. We first derive a new multi-step Value Iteration (VI) method that converges to the optimal Value Function (VF) with an exponential contraction rate but linear computational complexity. Given some trial, our so-called Highway RL performs rapid CA, by picking a policy and a possible lookahead (up to the trial end) that maximize the near-term reward during lookahead plus a DP-based estimate of the cumulative reward for the remaining part of the trial. Highway RL does not require off-policy corrections. Under mild assumptions, it achieves better convergence rates than the traditional one-step Bellman Optimality Operator. We then derive Highway Q-Learning, a convergent multi-step off-policy variant of Q-learning. We show that our Highway algorithms significantly outperform DP approaches on toy tasks. Finally, we propose a deep function approximation variant called Highway DQN. We evaluate it on visual MinAtar Games, outperforming similar multi-step methods.

****

Learning Locality and Isotropy in Dialogue Modeling

Han Wu,Haochen Tan,Mingjie Zhan,Gangming Zhao,Shaoqing Lu,Ding Liang,Linqi Song

Existing dialogue modeling methods have achieved promising performance on various dialogue tasks with the aid of Transformer and the large-scale pre-trained language models. However, some recent studies revealed that the context representations produced by these methods suffer the problem of anisotropy. In this paper, we find that the generated representations are also not conversational, losing the conversation structure information during the context modeling stage. To this end, we identify two properties in dialogue modeling, i.e., locality and isotropy, and present a simple method for dialogue representation calibration, namely SimDRC, to build isotropic and conversational feature spaces. Experimental results show that our approach significantly outperforms current state-of-the-art models on three open-domain dialogue tasks with eight benchmarks. More in-depth analyses further confirm the effectiveness of our proposed approach. We release the code at https://github.com/hahahawu/SimDRC.

****

Dynamic Historical Adaptation for Continual Image-Text Modeling

Yutian Luo,Yizhao Gao,Zhiwu Lu

In realistic application scenarios, existing methods for image-text modeling have limitations in dealing with data stream: training on all data needs too much computation/storage resources, and even the full access to previous data is invalid. In this work, we thus propose a new continual image-text modeling (CITM) setting that requires a model to be trained sequentially on a number of diverse image-text datasets. Although recent continual learning methods can be directly applied to the CITM setting, most of them only consider reusing part of previous data or aligning the output distributions of previous and new models, which is a partial or indirect way to acquire the old knowledge. In contrast, we propose a novel dynamic historical adaptation (DHA) method which can holistically and directly review the old knowledge from a historical model. Concretely, the historical model transfers its total parameters to the main/current model to utilize the holistic old knowledge. In turn, the main model dynamically transfers its parameters to the historical model at every five training steps to ensure that the knowledge gap between them is not too large. Extensive experiments show that our DHA outperforms other representative/latest continual learning methods under the CITM setting.

****

AutoGT: Automated Graph Transformer Architecture Search

Zizhao Zhang,Xin Wang,Chaoyu Guan,Ziwei Zhang,Haoyang Li,Wenwu Zhu

Although Transformer architectures have been successfully applied to graph data with the advent of Graph Transformer, current design of Graph Transformer still heavily relies on human labor and expertise knowledge to decide proper neural architectures and suitable graph encoding strategies at each Transformer layer. In literature, there have been some works on automated design of Transformers focusing on non-graph data such as texts and images without considering graph encoding strategies, which fail to handle the non-euclidean graph data. In this paper, we study the problem of automated graph Transformer, for the first time. However, solving these problems poses the following challenges: i) how can we design a unified search space for graph Transformer, and ii) how to deal with the coupling relations between Transformer architectures and the graph encodings of each Transformer layer. To address these challenges, we propose Automated Graph Transformer (AutoGT), a neural architecture search framework that can automatically discover the optimal graph Transformer architectures by joint optimization of Transformer architecture and graph encoding strategies. Specifically, we first propose a unified graph Transformer formulation that can represent most of state-of-the-art graph Transformer architectures. Based upon the unified formulation, we further design the graph Transformer search space that includes both candidate architectures and various graph encodings. To handle the coupling relations, we propose a novel encoding-aware performance estimation strategy by gradually training and splitting the supernets according to the correlations between graph encodings and architectures. The proposed strategy can provide a more consistent and fine-grained performance prediction when evaluating the jointly optimized graph encodings and architectures. Extensive experiments and ablation studies show that our proposed AutoGT gains sufficient improvement over state-of-the-art hand-crafted baselines on all datasets, demonstrating its effectiveness and wide applicability.

****

Cross Modal Domain Generalization for Query-based Video Segmentation

Yan Xia,Zhou Zhao,Jieming Zhu

Domain generalization (DG) aims to increase a model's generalization ability against the performance degradation when transferring to the target domains, which has been successfully applied in various visual and natural language tasks. However, DG on multi-modal tasks is still an untouched field. Compared with traditional single-modal DG, the biggest challenge of multi-modal DG is that each modality has to cope with its own domain shift. Directly applying the previous methods will make the generalization direction of the model in each modality inconsistent, resulting in negative effects when the model is migrated to the target domains. Thus in this paper, we explore the scenario of query-based video segmentation to study how to better advance the generalization ability of the model in the multi-modal situation. Considering the information from different modalities often shows consistency, we propose query-guided feature augmentation (QFA) and attention map adaptive instance normalization (AM-AdaIN) modules. Compared with traditional DG models, our method can combine visual and textual modalities together to guide each other for data augmentation and learn a domain-agnostic cross-modal relationship, which is more suitable for multi-modal transfer tasks. Extensive experiments on three query-based video segmentation generalization tasks demonstrate the effectiveness of our method.

****

Combating Exacerbated Heterogeneity for Robust Models in Federated Learning

Jianing Zhu,Jiangchao Yao,Tongliang Liu,quanming yao,Jianliang Xu,Bo Han

Privacy and security concerns in real-world applications have led to the development of adversarially robust federated models. However, the straightforward combination between adversarial training and federated learning in one framework can lead to the undesired robustness deterioration. We discover that the attribution behind this phenomenon is that the generated adversarial data could exacerbate the data heterogeneity among local clients, making the wrapped federated learning perform poorly. To deal with this problem, we propose a novel framework called Slack Federated Adversarial Training (SFAT), assigning the client-wise slack during aggregation to combat the intensified heterogeneity. Theoretically, we analyze the convergence of the proposed method to properly relax the objective when combining federated learning and adversarial training. Experimentally, we verify the rationality and effectiveness of SFAT on various benchmarked and real-world datasets with different adversarial training and federated optimization methods. The code is publicly available at: https://github.com/ZFancy/SFAT.

****

Pruning with Output Error Minimization for Producing Efficient Neural Networks

Koji Kamma,Toshikazu Wada

DNNs are known to have excessive parameters and thus are computationally expensive, which poses a challenge for implementations in various applications. Structured pruning is a technique of compressing a trained DNN model by removing redundant neurons (or channels). How well a pruned model maintains its accuracy depends on two factors. The first is compression ratio optimization, in other words, how many neurons are reduced in each layer. The other is layer-wise optimization, in other words, which neurons to be preserved in each layer. In this paper, we propose Pruning with Output Error Minimization (POEM), a layer-wise pruning method that conducts pruning and then performs reconstruction to compensate the error caused by pruning. The strength of POEM lies in its reconstruction using the weighted least squares method so as to minimize the output error of the activation function, while the previous methods minimize the error of the value before applying the activation function. The experiments with well-known DNN models and a large scale image recognition dataset show that POEM is better than the previous methods in maintaining the accuracy of those models.

****

Orientation-Aware Graph Neural Networks for Protein Structure Representation Learning

Jiahan Li,Shitong Luo,Congyue Deng,Chaoran Cheng,Jiaqi Guan,Leonidas Guibas,Jianzhu Ma,Jian Peng

By folding to particular 3D structures, proteins play a key role in living beings. To learn meaningful representation from a protein structure for downstream tasks, not only the global backbone topology but the local fine-grained orientational relations between amino acids should also be considered. In this work, we propose the Orientation-Aware Graph Neural Networks (OAGNNs) to better sense the geometric characteristics in protein structure (e.g. inner-residue torsion angles, inter-residue orientations). Extending a single weight from a scalar to a 3D vector, we construct a rich set of geometric-meaningful operations to process both the classical and SO(3) representations of a given structure. To plug our designed perceptron unit into existing Graph Neural Networks, we further introduce an equivariant message passing paradigm, showing superior versatility in maintaining SO(3)-equivariance at the global scale. Experiments have shown that our OAGNNs have a remarkable ability to sense geometric orientational features compared to classical networks. OAGNNs have also achieved state-of-the-art performance on various computational biology applications related to protein 3D structures.

****

Adaptive Update Direction Rectification for Unsupervised Continual Learning

Yizhao Gao,Nanyi Fei,Zhiwu Lu

Recent works on continual learning have shown that unsupervised continual learning (UCL) methods rival or even beat supervised continual learning methods. Howev

er, most UCL methods typically adopt fixed learning strategies with pre-defined objectives and ignore the influence of the constant shift of data distributions on the newer training process. This non-adaptive paradigm tends to achieve sub-optimal performance, since the optimal update direction (to ensure the trade-off between old and new tasks) keeps changing during training over sequential tasks. In this work, we thus propose a novel UCL framework termed AUDR to adaptively rectify the update direction by a policy network (i.e., the Actor) at each training step based on the reward predicted by a value network (i.e., the Critic). Concretely, different from existing Actor-Critic based reinforcement learning works, there are three vital designs that make our AUDR applicable to the UCL setting: (1) A reward function to measure the score/value of the currently selected action, which provides the ground-truth reward to guide the Critic's predictions; (2) An action space for the Actor to select actions (i.e., update directions) according to the reward predicted by the Critic; (3) A multinomial sampling strategy with a lower-bound on the sampling probability of each action, which is designed to improve the variance of the Actor's selected actions for more diversified exploration. Extensive experiments show that our AUDR achieves state-of-the-art results under both the in-dataset and cross-dataset UCL settings. Importantly, our AUDR also shows superior performance when combined with other UCL methods, which suggests that our AUDR is highly extensible and versatile.
****

Language Model Pre-training with Linguistically Motivated Curriculum Learning
Yile Wang,Yue Zhang,Peng Li,Yang Liu

Pre-training serves as a foundation of recent NLP models, where language modeling task is performed over large texts. It has been shown that data affects the quality of pre-training, and curriculum has been investigated regarding sequence length. We consider a linguistic perspective in the curriculum, where frequent words are learned first and rare words last. This is achieved by substituting syntactic constituents for rare words with their constituent labels. By such syntactic substitutions, a curriculum can be made by gradually introducing words with decreasing frequency levels. Without modifying model architectures or introducing external computational overhead, our data-centric method gives better performances over vanilla BERT on various downstream benchmarks.
****

Offline Reinforcement Learning with Closed-Form Policy Improvement Operators
Jiachen Li,Edwin Zhang,Ming Yin,Qinxun Bai,Yu-Xiang Wang,William Yang Wang

Behavior constrained policy optimization has been demonstrated to be a successful paradigm for tackling Offline Reinforcement Learning. By exploiting historical transitions, a policy is trained to maximize a learned value function while constrained by the behavior policy to avoid a significant distributional shift. In this paper, we propose our closed-form policy improvement operators. We make a novel observation that the behavior constraint naturally motivates the use of first-order Taylor approximation, leading to a linear approximation of the policy objective. Additionally, as practical datasets are usually collected by heterogeneous policies, we model the behavior policies as a Gaussian Mixture and overcome the induced optimization difficulties by leveraging the LogSumExp's lower bound and Jensen's Inequality, giving rise to a closed-form policy improvement operator. We instantiate an offline RL algorithm with our novel policy improvement operator and empirically demonstrate its effectiveness over state-of-the-art algorithms on the standard D4RL benchmark.
****

Sensitivity-aware Visual Parameter-efficient Tuning
Haoyu He,Jianfei Cai,Jing Zhang,Dacheng Tao,Bohan Zhuang

Visual Parameter-efficient Tuning (VPT) has become a powerful alternative for full fine-tuning, which only updates a small number of parameters while freezing the remaining vast majority of parameters to significantly reduce the storage costs for adapting the pre-trained vision models to downstream tasks. Although the storage burden is largely alleviated, VPT approaches still face many challenges, e.g., lower inference speed and lacking effective configurations for trainable parameters tailored for each task. In this paper, we present a simple yet effect ive approach termed Sensitivity-aware visual Parameter-efficient Tuning (SPT) to tackle these challenges. Given a desired tunable parameter budget, SPT quickly identifies the important parameters to the given task in a data-dependent way before fine-tuning, without the complex selection schedule. To increase the representational capacity at a negligible cost within the same parameter budget, we employ low-rank reparameterization to achieve a better trade-off between parameter efficiency and accuracy. Through extensive experiments on a wide range of downstream recognition tasks, our SPT achieves better overall transfer performance than the full fine-tuning and the other VPT approaches, with no additional computational or memory overhead during inference. For instance, SPT saves 99.35% of the trainable parameters than the full fine-tuning while achieving a 7.3% higher average top-1 accuracy on VTAB-1k benchmark with the supervised pre-trained ViT-B backbone. Notably, SPT is also the first work that bridges the gap between full fine-tuning and VPT approaches for backbones under self-supervised pre-training strategies MAE and MoCo v3 on the challenging VTAB-1k benchmark.

****

Towards Robust Object Detection Invariant to Real-World Domain Shifts

Qi Fan,Mattia Segu,Yu-Wing Tai,Fisher Yu,Chi-Keung Tang,Bernt Schiele,Dengxin Dai

Safety-critical applications such as autonomous driving require robust object detection invariant to real-world domain shifts. Such shifts can be regarded as different domain styles, which can vary substantially due to environment changes and sensor noises, but deep models only know the training domain style. Such domain style gap impedes object detection generalization on diverse real-world domains. Existing classification domain generalization (DG) methods cannot effectively solve the robust object detection problem, because they either rely on multiple source domains with large style variance or destroy the content structures of the original images. In this paper, we analyze and investigate effective solutions to overcome domain style overfitting for robust object detection without the above shortcomings. Our method, dubbed as Normalization Perturbation (NP), perturbs the channel statistics of source domain low-level features to synthesize various latent styles, so that the trained deep model can perceive diverse potential domains and generalizes well even without observations of target domain data in training. This approach is motivated by the observation that feature channel statistics of the target domain images deviate around the source domain statistics. We further explore the style-sensitive channels for effective style synthesis. Normalization Perturbation only relies on a single source domain and is surprisingly simple and effective, contributing a practical solution by effectively adapting or generalizing classification DG methods to robust object detection. Extensive experiments demonstrate the effectiveness of our method for generalizing object detectors under real-world domain shifts.

****

Light Sampling Field and BRDF Representation for Physically-based Neural Rendering

Jing Yang,Hanyuan Xiao,Wenbin Teng,Yunxuan Cai,Yajie Zhao

Physically-based rendering (PBR) is key for immersive rendering effects used widely in the industry to showcase detailed realistic scenes from computer graphics assets. A well-known caveat is that producing the same is computationally heavy and relies on complex capture devices. Inspired by the success in quality and efficiency of recent volumetric neural rendering, we want to develop a physically-based neural shader to eliminate device dependency and significantly boost performance. However, no existing lighting and material models in the current neural rendering approaches can accurately represent the comprehensive lighting models and BRDFs properties required by the PBR process. Thus, this paper proposes a novel lighting representation that models direct and indirect light locally through a light sampling strategy in a learned light sampling field. We also propose BRDF models to separately represent surface/subsurface scattering details to enable complex objects such as translucent material (i.e., skin, jade). We then implement our proposed representations with an end-to-end physically-based neural face skin shader, which takes a standard face asset (i.e., geometry, albedo map, and normal map) and an HDRI for illumination as inputs and generates a photo-realistic rendering as output. Extensive experiments showcase the quality and efficiency of our PBR face skin shader, indicating the effectiveness of our proposed lighting and material representations.

****

DREAM: Domain-free Reverse Engineering Attributes of Black-box Model

Rongqing Li,Jiaqi Yu,Changsheng Li,Wenhan Luo,Ye Yuan,Guoren Wang

Deep learning models are usually black boxes when deployed on machine learning platforms. Prior works have shown that the attributes (e.g., the number of convolutional layers) of a target black-box neural network can be exposed through a sequence of queries. There is a crucial limitation that these works assume the dataset used for training the target model to be known beforehand, and leverage this dataset for model attribute attack. However, it is difficult to access the training dataset of the target black-box model in reality. Therefore, whether the attributes of a target black-box model could be still revealed in this case is doubtful. In this paper, we investigate a new problem of Domain-free Reverse Engineering the Attributes of a black-box target Model, called DREAM, without requiring the model's training dataset available, and put forward a general and principled framework by casting this problem as an out of distribution (OOD) generalization problem. At the heart of our framework, we devise a multi-discriminator generative adversarial network (MDGAN) to learn domain invariant features. Based on these features, we can learn a domain-free model to inversely infer the attributes of a target black-box model with unknown training data. This makes our method one of the kinds that can gracefully apply to an arbitrary domain for model attribute reverse engineering with good generalization ability. Extensive experimental studies are conducted and the results validate the superiority of our proposed method over the baselines.

****

Structural Generalization of Visual Imitation Learning with Position-Invariant Regularization

Zhao-Heng Yin,Yang Gao,Qifeng Chen

How the visual imitation learning models can generalize to novel unseen visual observations is a highly challenging problem. Such a generalization ability is very crucial for their real-world applications. Since this generalization problem has many different aspects, we focus on one case called structural generalization, which refers to generalization to unseen task setup, such as a novel setup of object locations in the robotic manipulation problem. In this case, previous works observe that the visual imitation learning models will overfit to the absolute information (e.g., coordinates) rather than the relational information between objects, which is more important for decision making. As a result, the models will perform poorly in novel scenarios. Nevertheless, so far, it remains unclear how we can solve this problem effectively. Our insight into this problem is to explicitly remove the absolute information from the features learned by imitation learning models so that the models can use robust, relational information to make decisions. To this end, we propose a novel, position-invariant regularizer for generalization. The proposed regularizer will penalize the imitation learning model when its features contain absolute, positional information of objects. We carry out experiments on the MAGICAL and ProcGen benchmark, as well as a real-world robot manipulation problem. We find that our regularizer can effectively boost the structural generalization performance of imitation learning models. Through both qualitative and quantitative analysis, we verify that our method does learn robust relational representations.

****

Dealing with missing data using attention and latent space regularization

Jahan Che Penny-Dimri,Christoph Bergmeir,Julian Smith

Most practical data science problems encounter missing data. A wide variety of solutions exist, each with strengths and weaknesses that depend upon the missingness-generating process. Here we develop a theoretical framework for training and inference using only observed variables enabling modeling of incomplete dataset s without imputation. Using an information and measure-theoretic argument we construct models with latent space representations that regularize against the potential bias introduced by missing data. The theoretical properties of this approach are demonstrated empirically using a synthetic dataset. The performance of this approach is tested on 11 benchmarking datasets with missingness and 18 datasets corrupted across three missingness patterns with comparison against a state-of-the-art model and industry-standard imputation. We show that our proposed method outperforms common imputation methods and the current state-of-the-art with statistical significance.

****

Bidirectional Propagation for Cross-Modal 3D Object Detection

Yifan Zhang,Qijian Zhang,Junhui Hou,Yixuan Yuan,Guoliang Xing

Recent works have revealed the superiority of feature-level fusion for cross-modal 3D object detection, where fine-grained feature propagation from 2D image pixels to 3D LiDAR points has been widely adopted for performance improvement. Still, the potential of heterogeneous feature propagation between 2D and 3D domains has not been fully explored. In this paper, in contrast to existing pixel-to-point feature propagation, we investigate an opposite point-to-pixel direction, allowing point-wise features to flow inversely into the 2D image branch. Thus, when jointly optimizing the 2D and 3D streams, the gradients back-propagated from the 2D image branch can boost the representation ability of the 3D back-bone network working on LiDAR point clouds. Then, combining pixel-to-point and point-to-pixel information flow mechanisms, we further construct an interactive bidirectional feature propagation framework, dubbed BiProDet. In addition to the architectural design, we also propose normalized local coordinates map estimation, a new 2D auxiliary task for the training of the 2D image branch, which facilitates learning local spatial-aware features from the image modality and implicitly enhances the overall 3D detection performance. Extensive experiments and ablation studies validate the effectiveness of our method. Notably, we rank 1st on the highly competitive KITTI benchmark on the cyclist class by the time of submission. We also uploaded the source code in the supplementary material, which will be publicly available.

****

Policy Pre-training for Autonomous Driving via Self-supervised Geometric Modeling

Penghao Wu,Li Chen,Hongyang Li,Xiaosong Jia,Junchi Yan,Yu Qiao

Witnessing the impressive achievements of pre-training techniques on large-scale data in the field of computer vision and natural language processing, we wonder whether this idea could be adapted in a grab-and-go spirit, and mitigate the sample inefficiency problem for visuomotor driving. Given the highly dynamic and variant nature of the input, the visuomotor driving task inherently lacks the view and translation invariance, and the visual input contains massive irrelevant information for decision making, resulting in predominant pre-training approaches from general vision less suitable for the autonomous driving task. To this end, we propose PPGeo (Policy Pre-training via Geometric modeling), an intuitive and straightforward fully self-supervised framework curated for the policy pre-training in visuomotor driving. We aim at learning policy representations as a powerful abstraction by modeling 3D geometric scenes on large-scale unlabeled and uncalibrated YouTube driving videos. The proposed PPGeo is performed in two stages to support effective self-supervised training. In the first stage, the geometric modeling framework generates pose and depth predictions simultaneously, with two consecutive frames as input. In the second stage, the visual encoder learns driving policy representation by predicting the future ego-motion and optimizing with the photometric error based on current visual observation only. As such, the pre-trained visual encoder is equipped with rich driving policy related representations and thereby competent for multiple visuomotor driving tasks. As a side product, the pre-trained geometric modeling networks could bring further improvement to the depth and odometry estimation tasks. Extensive experiments covering a wide span of challenging scenarios have demonstrated the superiority of our pr oposed approach, where improvements range from 2% to even over 100% with very limited data.

****

Towards Expressive Graph Representations for Graph Neural Networks

Chengsheng Mao,Liang Yao,Yuan Luo

Graph Neural Network (GNN) shows its powerful capability for graph representation learning in various application areas. However, most existing GNN variants learn the graph representations in a non-injective or non-continuous fashion, both reducing the model expressive power. In this paper, we present a theoretical framework to improve the expressive power of GNN by taking both injectivity and continuity into account. Accordingly, we develop \textit{Injective Continuous Graph Neural Network} (ICGNN) that learns the graph and node representations in an injective and continuous fashion, so that it can map similar nodes or graphs to similar embeddings, and non-equivalent nodes or non-isomorphic graphs to different embeddings. We validate the proposed ICGNN model for graph classification and node classification on multiple benchmark datasets including both simple graphs and attributed graphs. The experimental results demonstrate that our model achieves state-of-the-art performances on most of the benchmarks.

****

EurNet: Efficient Multi-Range Relational Modeling of Spatial Multi-Relational Data

Minghao Xu,Yuanfan Guo,Yi Xu,Jian Tang,Xinlei Chen,Yuandong Tian

Modeling spatial relationship in the data remains critical across many different tasks, such as image classification, semantic segmentation and protein structure understanding. Previous works often use a unified solution like relative positional encoding. However, there exists different kinds of spatial relations, including short-range, medium-range and long-range relations, and modeling them separately can better capture the focus of different tasks on the multi-range relations (e.g., short-range relations can be important in instance segmentation, while long-range relations should be upweighted for semantic segmentation). In this work, we introduce the EurNet for Efficient multi-range relational modeling. EurNet constructs the multi-relational graph, where each type of edge corresponds to short-, medium- or long-range spatial interactions. In the constructed graph, EurNet adopts a novel modeling layer, called gated relational message passing (GRMP), to propagate multi-relational information across the data. GRMP captures multiple relations within the data with little extra computational cost. We study EurNets in two important domains for image and protein structure modeling. Extensive experiments on ImageNet classification, COCO object detection and ADE20K semantic segmentation verify the gains of EurNet over the previous SoTA FocalNet. On the EC and GO protein function prediction benchmarks, EurNet consistently surpasses the previous SoTA GearNet. Our results demonstrate the strength of EurNets on modeling spatial multi-relational data from various domains.

****

TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis

Haixu Wu,Tengge Hu,Yong Liu,Hang Zhou,Jianmin Wang,Mingsheng Long

Time series analysis is of immense importance in extensive applications, such as weather forecasting, anomaly detection, and action recognition. This paper focuses on temporal variation modeling, which is the common key problem of extensive analysis tasks. Previous methods attempt to accomplish this directly from the 1D time series, which is extremely challenging due to the intricate temporal patterns. Based on the observation of multi-periodicity in time series, we ravel out the complex temporal variations into the multiple intraperiod- and interperiod-variations. To tackle the limitations of 1D time series in representation capability, we extend the analysis of temporal variations into the 2D space by transforming the 1D time series into a set of 2D tensors based on multiple periods. This transformation can embed the intraperiod- and interperiod-variations into the columns and rows of the 2D tensors respectively, making the 2D-variations to be easily modeled by 2D kernels. Technically, we propose the TimesNet with TimesBlock as a task-general backbone for time series analysis. TimesBlock can discover the multi-periodicity adaptively and extract the complex temporal variations fro m transformed 2D tensors by a parameter-efficient inception block. Our proposed TimesNet achieves consistent state-of-the-art in five mainstream time series analysis tasks, including short- and long-term forecasting, imputation, classification, and anomaly detection. Code is available at this repository: https://github.com/thuml/TimesNet.

****

Learning without Prejudices: Continual Unbiased Learning via Benign and Malignant Forgetting

Myeongho Jeon,Hyoje Lee,Yedarm Seong,Myungjoo Kang

Although machine learning algorithms have achieved state-of-the-art status in image classification, recent studies have substantiated that the ability of the models to learn several tasks in sequence, termed continual learning (CL), often suffers from abrupt degradation of performance from previous tasks. A large body of CL frameworks has been devoted to alleviating this issue. However, we observe that forgetting phenomena in CL are not always unfavorable, especially when there is bias (spurious correlation) in training data. We term such type of forgetting benign forgetting, and categorize detrimental forgetting as malignant forgetting. Based on this finding, our objective in this study is twofold: (a) to discourage malignant forgetting by generating previous representations, and (b) encourage benign forgetting by employing contrastive learning in conjunction with feature-level augmentation. Extensive evaluations of biased experimental setups demonstrate that our proposed method, Learning without Prejudices, is effective for continual unbiased learning.

****

FINDE: Neural Differential Equations for Finding and Preserving Invariant Quantities

Takashi Matsubara,Takaharu Yaguchi

Many real-world dynamical systems are associated with first integrals (a.k.a. invariant quantities), which are quantities that remain unchanged over time. The discovery and understanding of first integrals are fundamental and important topics both in the natural sciences and in industrial applications. First integrals arise from the conservation laws of system energy, momentum, and mass, and from constraints on states; these are typically related to specific geometric structures of the governing equations. Existing neural networks designed to ensure such first integrals have shown excellent accuracy in modeling from data. However, these models incorporate the underlying structures, and in most situations where neural networks learn unknown systems, these structures are also unknown. This limitation needs to be overcome for scientific discovery and modeling of unknown systems. To this end, we propose first integral-preserving neural differential equation (FINDE). By leveraging the projection method and the discrete gradient method, FINDE finds and preserves first integrals from data, even in the absence of prior knowledge about underlying structures. Experimental results demonstrate that FINDE can predict future states of target systems much longer and find various quantities consistent with well-known first integrals in a unified manner.

****

Approximate Vanishing Ideal Computations at Scale

Elias Samuel Wirth,Hiroshi Kera,Sebastian Pokutta

The vanishing ideal of a set of points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}\subseteq \mathbb{R}^n$ is the set of polynomials that evaluate to 0 over all points $\mathbf{x} \in X$ and admits an efficient representation by a finite subset of generators. In practice, to accommodate noise in the data, algorithms that construct generators of the approximate vanishing ideal are widely studied but their computational complexities remain expensive. In this paper, we scale up the oracle approximate vanishing ideal algorithm (OAVI), the only generator-constructing algorithm with known learning guarantees. We prove that the computational complexity of OAVI is not superlinear, as previously claimed, but linear in the number of samples m. In addition, we propose two modifications that accelerate OAVI's training time: Our analysis reveals that replacing the pairwise conditional gradients algorithm, one of the solvers used in OAVI, with the faster blended pairwise conditional gradients algorithm leads to an exponential speed-up in the number of features n. Finally, using a new inverse Hessian boosting approach, intermediate convex optimization problems can be solved almost instantly, improving OAVI's training time by multiple orders of magnitude in a variety of numerical experiments.

****

Understanding Incremental Learning of Gradient Descent: A Fine-grained analysis of Matrix Sensing

Jikai Jin,Zhiyuan Li,Kaifeng Lyu,Simon Shaolei Du,Jason D. Lee

The implicit bias of optimization algorithms such as gradient descent (GD) is believed to play an important role in generalization of modern machine learning methods such as deep learning. This paper provides a fine-grained analysis of the dynamics of GD for the matrix sensing problem, whose goal is to recover a low-rank ground-truth matrix from near-isotropic linear measurements. With small initialization, we that GD behaves similarly to the greedy low-rank learning heuristics~\citep{li2020towards} and follows an incremental learning procedure~\citep{gissin2019implicit}. That is, GD sequentially learns solutions with increasing ranks until it recovers the ground-truth matrix. Compared to existing works which only analyze the first learning phase for rank-1 solutions, our result is stronger because it characterizes the whole learning process. Moreover, our analysis of the incremental learning procedure applies to the
under-parameterized regime as well. As a key ingredient of our analysis, we observe that GD always follows an approximately low-rank trajectory and develops novel landscape properties for matrix sensing with low-rank parameterization. Finally, we conduct numerical experiments which confirm our theoretical findings.

****

Selective Annotation Makes Language Models Better Few-Shot Learners

Hongjin SU,Jungo Kasai,Chen Henry Wu,Weijia Shi,Tianlu Wang,Jiayi Xin,Rui Zhang,Mari Ostendorf,Luke Zettlemoyer,Noah A. Smith,Tao Yu

Many recent approaches to natural language tasks are built on the remarkable abilities of large language models. Large language models can perform in-context learning, where they learn a new task from a few task demonstrations, without any parameter updates. This work examines the implications of in-context learning for the creation of datasets for new natural language tasks. Departing from recent in-context learning methods, we formulate an annotation-efficient, two-step framework: selective annotation that chooses a pool of examples to annotate from unlabeled data in advance, followed by prompt retrieval that retrieves task examples from the annotated pool at test time. Based on this framework, we propose an unsupervised, graph-based selective annotation method, voke-k, to select diverse, representative examples to annotate. Extensive experiments on 10 datasets (covering classification, commonsense reasoning, dialogue, and text/code generation) demonstrate that our selective annotation method improves the task performance by a large margin. On average, vote-k achieves a 12.9%/11.4% relative gain under an annotation budget of 18/100, as compared to randomly selecting examples to annotate. Compared to state-of-the-art supervised finetuning approaches, it yields similar performance with 10-100x less annotation cost across 10 tasks. We further analyze the effectiveness of our framework in various scenarios: language models with varying sizes, alternative selective annotation methods, and cases where there is a test data domain shift. We hope that our studies will serve as a basis for data annotations as large language models are increasingly applied to new tasks.

****

Switch-NeRF: Learning Scene Decomposition with Mixture of Experts for Large-scale Neural Radiance Fields

Zhenxing MI,Dan Xu

The Neural Radiance Fields (NeRF) have been recently applied to reconstruct building-scale and even city-scale scenes. To model a large-scale scene efficiently, a dominant strategy is to employ a divide-and-conquer paradigm via performing scene decomposition, which decomposes a complex scene into parts that are further processed by different sub-networks. Existing large-scale NeRFs mainly use heuristic hand-crafted scene decomposition, with regular 3D-distance-based or physic al-street-block-based schemes. Although achieving promising results, the hand-crafted schemes limit the capabilities of NeRF in large-scale scene modeling in several aspects. Manually designing a universal scene decomposition rule for different complex scenes is challenging, leading to adaptation issues for different scenarios. The decomposition procedure is not learnable, hindering the network from jointly optimizing the scene decomposition and the radiance fields in an end-to-end manner. The different sub-networks are typically optimized independently, and thus hand-crafted rules are required to composite them to achieve a better consistency. To tackle these issues, we propose Switch-NeRF, a novel end-to-end large-scale NeRF with learning-based scene decomposition. We design a gating network to dispatch 3D points to different NeRF sub-networks. The gating network can be optimized together with the NeRF sub-networks for different scene partitions, by a design with the Sparsely Gated Mixture of Experts (MoE). The outputs from different sub-networks can also be fused in a learnable way in the unified framework to effectively guarantee the consistency of the whole scene. Furthermore, the proposed MoE-based Switch-NeRF model is carefully implemented and optimized to achieve both high-fidelity scene reconstruction and efficient computation. Our method establishes clear state-of-the-art performances on several large-scale datasets. To the best of our knowledge, we are the first to propose an applicable end-to-end sparse NeRF network with learning-based decomposition for large-scale scenes. Codes are released at https://github.com/MiZhenxing/Switch-NeRF.

****

Efficient, Stable, and Analytic Differentiation of the Sinkhorn Loss
Yixuan Qiu,Haoyun Yin,Xiao Wang

Optimal transport and the Wasserstein distance have become indispensable building blocks of modern deep generative models, but their computional costs greatly prohibit their applications in statistical machine learning models. Recently, the Sinkhorn loss, as an approximation to the Wasserstein distance, has gained massive popularity, and much work has been done for its theoretical properties. To embed the Sinkhorn loss into gradient-based learning frameworks, efficient algorithms for both the forward and backward passes of the Sinkhorn loss are required. In this article, we first demonstrate issues of the widely-used Sinkhorn's algorithm, and show that the L-BFGS algorithm is a potentially better candidate for the forward pass. Then we derive an analytic form of the derivative of the Sinkhorn loss with respect to the input cost matrix, which results in a very efficient backward algorithm. We rigorously analyze the convergence and stability properties of the advocated algorithms, and use various numerical experiments to validate the superior performance of the proposed methods.

****

A Holistic View of Label Noise Transition Matrix in Deep Learning and Beyond
LIN Yong,Renjie Pi,WEIZHONG ZHANG,Xiaobo Xia,Jiahui Gao,Xiao Zhou,Tongliang Liu,Bo Han

In this paper, we explore learning statistically consistent classifiers under label noise by estimating the noise transition matrix T. We first provide a holistic view of existing T-estimation methods including those with or without anchor point assumptions. We unified them into the Minimum Geometric Envelope Operator (MGEO) framework, which tries to find the smallest T (in terms of a certain metric) that elicits a convex hull to enclose the posteriors of all the training data. Although MGEO methods show appealing theoretical properties and empirical results, we find them prone to failing when the noisy posterior estimation is imperfect, which is inevitable in practice. Specifically, we show that MGEO methods are in-consistent even with infinite samples if the noisy posterior is not estimated accurately. In view of this, we make the first effort to address this issue by proposing a novel T-estimation framework via the lens of bilevel optimization, and term it RObust Bilevel OpTimzation (ROBOT). ROBOT paves a new road beyond MGEO framework, which enjoys strong theoretical properties: identifibility, consistency and finite-sample generalization guarantees. Notably, ROBOT neither requires the perfect posterior estimation nor assumes the existence of anchor points. We further theoretically demonstrate that ROBOT is more robust in the case where MGEO methods fail. Experimentally, our framework also shows superior perform ance across multiple benchmarks.

****

Active Learning in Bayesian Neural Networks with Balanced Entropy Learning Principle

Jae Oh Woo

Acquiring labeled data is challenging in many machine learning applications with limited budgets. Active learning gives a procedure to select the most informative data points and improve data efficiency by reducing the cost of labeling. The info-max learning principle maximizing mutual information such as BALD has been successful and widely adapted in various active learning applications. However, this pool-based specific objective inherently introduces a redundant selection and further requires a high computational cost for batch selection. In this paper, we design and propose a new uncertainty measure, Balanced Entropy Acquisition (BalEntAcq), which captures the information balance between the uncertainty of underlying softmax probability and the label variable. To do this, we approximate each marginal distribution by Beta distribution. Beta approximation enables us to formulate BalEntAcq as a ratio between an augmented entropy and the marginalized joint entropy. The closed-form expression of BalEntAcq facilitates parallelization by estimating two parameters in each marginal Beta distribution. BalEntAcq is a purely standalone measure without requiring any relational computations with other data points. Nevertheless, BalEntAcq captures a well-diversified selection near the decision boundary with a margin, unlike other existing uncertainty measures such as BALD, Entropy, or Mean Standard Deviation (MeanSD). Finally, we demonstrate that our balanced entropy learning principle with BalEntAcq consistently outperforms well-known linearly scalable active learning methods, including a recently proposed PowerBALD, a simple but diversified version of BALD, by showing experimental results obtained from MNIST, CIFAR-100, SVHN, and TinyImageNet datasets.

****

Near-Optimal Adversarial Reinforcement Learning with Switching Costs

Ming Shi,Yingbin Liang,Ness Shroff

Switching costs, which capture the costs for changing policies, are regarded as a critical metric in reinforcement learning (RL), in addition to the standard metric of losses (or rewards). However, existing studies on switching costs (with a coefficient that is strictly positive and is independent of the time horizon) have mainly focused on static RL, where the loss distribution is assumed to be fixed during the learning process, and thus practical scenarios where the loss distribution could be non-stationary or even adversarial are not considered. While adversarial RL better models this type of practical scenarios, an open problem remains: how to develop a provably efficient algorithm for adversarial RL with switching costs? This paper makes the first effort towards solving this problem. First, we provide a regret lower-bound that shows that the regret of any algorithm must be larger than $\tilde{\Omega}((H S A)^{1/3} T^{2/3})$, where $T$, $S$, $A$ and $H$ are the number of episodes, states, actions and layers in each episode, respectively. Our lower bound indicates that, due to the fundamental challenge of switching costs in adversarial RL, the best achieved regret (whose dependency on $T$ is $\tilde{O}(\sqrt{T})$) in static RL with switching costs (as well as adversarial RL without switching costs) is no longer achievable. Moreover, we propose two novel switching-reduced algorithms with regrets that match our lower bound when the transition function is known, and match our lower bound within a small factor of $\tilde{O}(H^{1/3})$ when the transition function is unknown. Our regret analysis demonstrates the near-optimal performance of them.

****

NORM: Knowledge Distillation via N-to-One Representation Matching

Xiaolong Liu,Lujun Li,Chao Li,Anbang Yao

Existing feature distillation methods commonly adopt the One-to-one Representation Matching between any pre-selected teacher-student layer pair. In this paper, we present $N$-to-$O$ne $R$epresentation $M$atching (NORM), a new two-stage knowledge distillation method, which relies on a simpleFeature Transform (FT) module consisting of two linear layers. In view of preserving the intact information l earnt by the teacher network, during training, our FT module is merely inserted after the last convolutional layer of the student network. The first linear layer projects the student representation to a feature space having N times feature channels than the teacher representation from the last convolutional layer, and the second linear layer contracts the expanded output back to the original feature space. By sequentially splitting the expanded student representation into N non-overlapping feature segments having the same number of feature channels as the teacher's, they can be readily forced to approximate the intact teacher representation simultaneously, formulating a novel many-to-one representation matching mechanism conditioned on a single teacher-student layer pair. After training, such an FT module will be naturally merged into the subsequent fully connected layer thanks to its linear property, introducing no extra parameters or architectural modifications to the student network at inference. Extensive experiments on different visual recognition benchmarks demonstrate the leading performance of our method. For instance, the ResNet18|MobileNet|ResNet50-1/4 model trained by NORM reaches 72.14%|74.26%|68.03% top-1 accuracy on the ImageNet dataset when using a pre-trained ResNet34|ResNet50|ResNet50 model as the teacher, achieving an absolute improvement of 2.01%|4.63%|3.03% against the individually trained counterpart. Code is available at https://github.com/OSVAI/NORM.

****

Downstream Datasets Make Surprisingly Good Pretraining Corpora

Kundan Krishna,Saurabh Garg,Jeffrey P. Bigham,Zachary Chase Lipton

For most natural language processing tasks, the dominant practice is to finetune large pretrained transformer models (e.g., BERT) using smaller downstream datasets. Despite the success of this approach, it remains unclear to what extent these gains are attributable to the massive background corpora employed for pretraining versus to the pretraining objectives themselves. This paper introduces a large-scale study of self-pretraining, where the same (downstream) training data is used for both pretraining and finetuning. In experiments addressing both ELECTRA and RoBERTa models and 10 distinct downstream datasets, we observe that self-pretraining rivals standard pretraining on the BookWiki corpus (despite using around $10\times$--$500\times$ less data), outperforming the latter on 7 and 5 datasets, respectively. Surprisingly, these task-specific pretrained models often perform well on other tasks, including the GLUE benchmark. Our results suggest that in many scenarios, performance gains attributable to pretraining are driven primarily by the pretraining objective itself and are not always attributable to the incorporation of massive datasets. These findings are especially relevant in light of concerns about intellectual property and offensive content in web-scale pretraining data.

****

Revisiting Embeddings for Graph Neural Networks

Skye Purchase,Yiren Zhao,Robert D. Mullins

Current graph representation learning techniques use Graph Neural Networks (GNNs) to extract features from dataset embeddings. In this work, we examine the quality of these embeddings and assess how changing them can affect the accuracy of GNNs. We explore different embedding extraction techniques for both images and texts; and find that the choice of embedding biases the performance of different GNN architectures and thus the choice of embedding influences the selection of GNNs regardless of the underlying dataset. In addition, we only see an improvement in accuracy from some GNN models compared to the accuracy of models trained from scratch or fine-tuned on the underlying data without utilising the graph connections. As an alternative, we propose Graph-connected Network (GraNet) layers to better leverage existing unconnected models within a GNN. Existing language and vision models are thus improved by allowing neighbourhood aggregation. This gives a chance for the model to use pre-trained weights, if possible, and we demonstrate that this approach improves the accuracy compared to traditional GNNs: on Flickr v2, GraNet beats GAT2 and GraphSAGE by 7.7% and 1.7% respectively.

****

Empirical analysis of representation learning and exploration in neural kernel bandits

Michal Lisicki,Arash Afkanpour,Graham W. Taylor

Neural bandits have been shown to provide an efficient solution to practical sequential decision tasks that have nonlinear reward functions. The main contributor to that success is approximate Bayesian inference, which enables neural network (NN) training with uncertainty estimates. However, Bayesian NNs often suffer from a prohibitive computational overhead or operate on a subset of parameters. Alternatively, certain classes of infinite neural networks were shown to directly correspond to Gausian processes (GP) with neural kernels (NK). NK-GPs provide accurate uncertainty estimates and can be trained faster than most Bayesian NNs. We propose to guide common bandit policies with NK distributions and show that NK bandits achieve state-of-the-art performance on nonlinear structured data. Moreover, we propose a framework for measuring independently the ability of a bandit algorithm to learn representations and explore, and use it to analyze the impact of NK distributions w.r.t. those two aspects. We consider policies based on a GP and a Student's t-process (TP). Furthermore, we study practical considerations, such as training frequency and model partitioning. We believe our work will help better understand the impact of utilizing NKs in applied settings.

****

Exploiting Spatial Separability for Deep Learning Multichannel Speech Enhancement with an Align-and-Filter Network

Ching-Hua Lee,Chouchang Yang,Yilin Shen,Hongxia Jin

Multichannel speech enhancement (SE) systems separate the target speech from background noise by performing spatial and spectral filtering. The development of multichannel SE has a long history in the signal processing field, where one crucial step is to exploit spatial separability of sound sources by aligning the microphone signals in response to the target speech source prior to further filtering processes. However, most existing deep learning based multichannel SE works have yet to effectively incorporate or emphasize this spatial alignment aspect in the network design – we postulate that it is owing to the lack of suitable datasets with sufficient spatial diversity of the speech sources. In this paper, we highlight this important but often overlooked step in deep learning based multichannel SE, i.e., signal alignment, by introducing an Align-and-Filter network (AFnet) featuring a two-stage sequential masking design. The AFnet estimates two sets of masks, the alignment masks and filtering masks, and multiplies the estimated masks with the respective input signals to each stage sequentially, while leveraging the relative transfer functions (RTFs) for guiding the model to align signals with various speech source locations during training. For exploration purposes, we argue that the popular CHiME-3 multichannel dataset has its own limitation in representing spatially diverse speech data as the speakers were mostly located at the front side, and thereby adopt simulated and real-world measured room impulse responses to generate multichannel recordings where the target speech sources might come from arbitrary directions. Our findings suggest that for spatially diverse speaker scenarios, careful consideration of exploiting spatial characteristics is of great importance for deep learning based multichannel SE especially when the number of microphone gets increased. We show that utilizing the RTFs for signal alignment purposes in the two-stage, sequential masking framework consistently improves the capability of the network to separate the target speech from the noise signals, supporting that spatial separability is being effectively exploited by the proposed model. Our studies advocate for the advantages and significance of considering the signal alignment aspect, a wisdom coming from conventional signal processing, for developing future deep based multichannel SE algorithms to improve enhancement outcomes with positional diverse target speech scenarios.

****

CroMA: Cross-Modality Adaptation for Monocular BEV Perception

Yunze Man,Liangyan Gui,Yu-Xiong Wang

Incorporating multiple sensor modalities, and closing the domain gaps between tr aining and deployment are two challenging yet critical topics for self-driving. Existing adaption works only focus on visual-level domain gaps, overlooking the sensor-type gaps which exist in reality. A model trained with a collection of sensor modalities may need to run on another setting with less types of sensors available. In this work, we propose a Cross-Modality Adaptation (CroMA) framework to facilitate the learning of a more robust monocular BEV perception model, which transfer the point clouds knowledge from LiDAR sensor during training phase to the camera-only testing scenario. The absence of LiDAR during testing negates the usage of it as model input. Hence, our key idea lies in the design of a LiDAR-teacher and Camera-student knowledge distillation model, as well as a multi-level adversarial learning mechanism, which adapt and align the features learned from different sensors and domains. This work results in the first open analysis of cross-domain perception and cross-sensor adaptation model for monocular 3D tasks in the wild. We benchmark our approach on large-scale datasets under various domain shifts and show state-of-the-art results against various baselines.

****

CausalAgents: A Robustness Benchmark for Motion Forecasting Using Causal Relationships

Rebecca Roelofs,Liting Sun,Benjamin Caine,Khaled S. Refaat,Benjamin Sapp,Scott Ettinger,Wei Chai

As machine learning models become increasingly prevalent in motion forecasting systems for autonomous vehicles (AVs), it is critical that we ensure that model predictions are safe and reliable. However, exhaustively collecting and labeling the data necessary to fully test the long tail of rare and challenging scenarios is difficult and expensive. In this work, we construct a new benchmark for evaluating and improving model robustness by applying perturbations to existing data. Specifically, we conduct an extensive labeling effort to identify causal agents, or agents whose presence influences human driver behavior in any way, in the Waymo Open Motion Dataset (WOMD), and we use these labels to perturb the data by deleting non-causal agents from the scene. We then evaluate a diverse set of state-of-the-art deep-learning model architectures on our proposed benchmark and find that all models exhibit large shifts under perturbation. Under non-causal perturbations, we observe a 25-38% relative change in minADE as compared to the original. We then investigate techniques to improve model robustness, including increasing the training dataset size and using targeted data augmentations that drop agents throughout training. We provide the causal agent labels as an additional attribute to WOMD and release the robustness benchmarks to aid the community in building more reliable and safe deep-learning models for motion forecasting.

****

Adaptive Client Sampling in Federated Learning via Online Learning with Bandit Feedback

Boxin Zhao,Ziqi Liu,Chaochao Chen,mladen kolar,Zhiqiang Zhang,JUN ZHOU

Due to the high cost of communication, federated learning (FL) systems need to sample a subset of clients that are involved in each round of training. As a result, client sampling plays an important role in FL systems as it affects the convergence rate of optimization algorithms used to train machine learning models. Despite its importance, there is limited work on how to sample clients effectively. In this paper, we cast client sampling as an online learning task with bandit feedback, which we solve with an online stochastic mirror descent (OSMD) algorithm designed to minimize the sampling variance. We then theoretically show how our sampling method can improve the convergence speed of optimization algorithms. To handle the tuning parameters in OSMD that depend on the unknown problem parameters, we use the online ensemble method and doubling trick. We prove a dynamic regret bound relative to any sampling sequence. The regret bound depends on the total variation of the comparator sequence, which naturally captures the intrinsic difficulty of the problem. To the best of our knowledge, these theoretical contributions are new and the proof technique is of independent interest. Through both synthetic and real data experiments, we illustrate advantages of the proposed client sampling algorithm over the widely used uniform sampling and existing online learning based sampling strategies. The proposed adaptive sampling procedure is applicable beyond the FL problem studied here and can be used to improve the performance of stochastic optimization procedures such as stochastic gradient descent and stochastic coordinate descent.

****

Dynamical Isometry for Residual Networks

Advait Harshal Gadhikar,Rebekka Burkholz

The training success, training speed and generalization ability of neural networks rely crucially on the choice of random parameter initialization. It has been shown for multiple architectures that initial dynamical isometry is particularly advantageous. Known initialization schemes for residual blocks, however, miss this property and suffer from degrading separability of different inputs for increasing depth and instability without Batch Normalization or lack feature diversity. We propose a random initialization scheme, Risotto, that achieves perfect dynamical isometry for residual networks with ReLU activation functions even for finite depth and width. It balances the contributions of the residual and skip branches unlike other schemes, which initially bias towards the skip connections. In experiments, we demonstrate that in most cases our approach outperforms initialization schemes proposed to make Batch Normalization obsolete, including Fixup and SkipInit, and facilitates stable training. Also in combination with Batch Normalization, we find that Risotto often achieves the overall best result.

****

How Erdös and Rényi Win the Lottery

Advait Harshal Gadhikar,Sohom Mukherjee,Rebekka Burkholz

Random masks define surprisingly effective sparse neural network models, as has been shown empirically. The resulting Erd\"os-R\'enyi (ER) random graphs can often compete with dense architectures and state-of-the-art lottery ticket pruning algorithms struggle to outperform them, even though the random baselines do not rely on computationally expensive pruning-training iterations but can be drawn initially without significant computational overhead. We offer a theoretical explanation of how such ER masks can approximate arbitrary target networks if they are wider by a logarithmic factor in the inverse sparsity $1 / \log(1/\text{sparsity})$. While we are the first to show theoretically and experimentally that random ER source networks contain strong lottery tickets, we also prove the existence of weak lottery tickets that require a lower degree of overparametrization than strong lottery tickets. These unusual results are based on the observation that ER masks are well trainable in practice, which we verify in experiments with varied choices of random masks. Some of these data-free choices outperform previously proposed random approaches on standard image classification benchmark datasets.

****

GPViT: A High Resolution Non-Hierarchical Vision Transformer with Group Propagation

Chenhongyi Yang,Jiarui Xu,Shalini De Mello,Elliot J. Crowley,Xiaolong Wang

We present the Group Propagation Vision Transformer (GPViT): a novel non- hierarchical (i.e. non-pyramidal) transformer model designed for general visual recognition with high-resolution features. High-resolution features (or tokens) are a natural fit for tasks that involve perceiving fine-grained details such as detection and segmentation, but exchanging global information between these features is expensive in memory and computation because of the way self-attention scales. We provide a highly efficient alternative Group Propagation Block (GP Block) to exchange global information. In each GP Block, features are first grouped to- gether by a fixed number of learnable group tokens; we then perform Group Propagation where global information is exchanged between the grouped fea- tures; finally, global information in the updated grouped features is returned back to the image features through a transformer decoder. We evaluate GPViT on a variety of visual recognition tasks including image classification, semantic seg- mentation, object detection, and instance segmentation. Our method achieves significant performance gains over previous works across all tasks, especially on tasks that require high-resolution outputs, for example, our GPViT-L3 out- performs Swin Tra nsformer-B by 2.0 mIoU on ADE20K semantic segmentation with only half as many parameters. Code and pre-trained models are available at https://github.com/ChenhongyiYang/GPViT.

****

Variational Imbalanced Regression

Ziyan Wang,Hao Wang

Existing regression models tend to fall short in both accuracy and uncertainty estimation when the label distribution is imbalanced. In this paper, we propose a probabilistic deep learning model, dubbed variational imbalanced regression (VIR), which not only performs well in imbalanced regression but naturally produces reasonable uncertainty estimation as a byproduct. Different from typical variational autoencoders assuming I.I.D. representation (a data point's representation is not directly affected by other data points), our VIR borrows data with similar regression labels to compute the latent representation's variational distribution; furthermore, different from deterministic regression models producing point estimates, VIR predicts the entire normal-inverse-gamma distributions and modulates the associated conjugate distributions to impose probabilistic reweighting on the imbalanced data, thereby providing better uncertainty estimation. Experiments in several real-world datasets show that our VIR can outperform state-of-the-art imbalanced regression models in terms of both accuracy and uncertainty estimation.

****

Critic Sequential Monte Carlo

Vasileios Lioutas,Jonathan Wilder Lavington,Justice Sefas,Matthew Niedoba,Yunpeng Liu,Berend Zwartsenberg,Setareh Dabiri,Frank Wood,Adam Scibior

We introduce CriticSMC, a new algorithm for planning as inference built from a composition of sequential Monte Carlo with learned Soft-Q function heuristic factors. These heuristic factors, obtained from parametric approximations of the marginal likelihood ahead, more effectively guide SMC towards the desired target distribution, which is particularly helpful for planning in environments with hard constraints placed sparsely in time. Compared with previous work, we modify the placement of such heuristic factors, which allows us to cheaply propose and evaluate large numbers of putative action particles, greatly increasing inference and planning efficiency. CriticSMC is compatible with informative priors, whose density function need not be known, and can be used as a model-free control algorithm. Our experiments on collision avoidance in a high-dimensional simulated driving task show that CriticSMC significantly reduces collision rates at a low computational cost while maintaining realism and diversity of driving behaviors across vehicles and environment scenarios.

****

Radial Spike and Slab Bayesian Neural Networks for Sparse Data in Ransomware Attacks

Jurijs Nazarovs,Jack W Stokes,Melissa Turcotte,Justin Carroll,Itai Grady

Ransomware attacks are increasing at an alarming rate, leading to large financial losses, unrecov- erable encrypted data, data leakage, and privacy concerns. The prompt detection of ransomware attacks is required to minimize further damage, particularly during the encryption stage. However, the frequency and structure of the observed ransomware attack data makes this task difficult to accomplish in practice. The data corresponding to ransomware attacks represents temporal, high- dimensional sparse signals, with limited records and very imbalanced classes. While traditional deep learning models have been able to achieve state-of-the-art results in a wide variety of domains, Bayesian Neural Networks, which are a class of probabilistic models, are better suited to the issues of the ransomware data. These models combine ideas from Bayesian statistics with the rich expres- sive power of neural networks. In this paper, we propose the Radial Spike and Slab Bayesian Neural Network, which is a new type of Bayesian Neural network that includes a new form of the approx- imate posterior distribution. The model scales well to large architectures and recovers the sparse structure of target functions. We provide a theoretical justification for using this type of distribution, as well as a computationally efficient method to perform variational inference

. We demonstrate the performance of our model on a real dataset of ransomware attacks and show improvement over a large number of baselines, including state-of-the-art models such as Neural ODEs (ordinary dif- ferential equations). In addition, we propose to represent low-level events as MITRE ATT&CK tactics, techniques, and procedures (TTPs) which allows the model to better generalize to unseen ransomware attacks.

****

Autoencoders as Cross-Modal Teachers: Can Pretrained 2D Image Transformers Help 3D Representation Learning?

Runpei Dong,Zekun Qi,Linfeng Zhang,Junbo Zhang,Jianjian Sun,Zheng Ge,Li Yi,Kaisheng Ma

The success of deep learning heavily relies on large-scale data with comprehensive labels, which is more expensive and time-consuming to fetch in 3D compared to 2D images or natural languages. This promotes the potential of utilizing models pretrained with data more than 3D as teachers for cross-modal knowledge transferring. In this paper, we revisit masked modeling in a unified fashion of knowledge distillation, and we show that foundational Transformers pretrained with 2D images or natural languages can help self-supervised 3D representation learning through training Autoencoders as Cross-Modal Teachers (ACT). The pretrained Transformers are transferred as cross-modal 3D teachers using discrete variational autoencoding self-supervision, during which the Transformers are frozen with prompt tuning for better knowledge inheritance. The latent features encoded by the 3D teachers are used as the target of masked point modeling, wherein the dark knowledge is distilled to the 3D Transformer students as foundational geometry understanding. Our ACT pretrained 3D learner achieves state-of-the-art generalization capacity across various downstream benchmarks, e.g., 88.21% overall accuracy on ScanObjectNN. Codes have been released at https://github.com/RunpeiDong/ACT.

****

Explainability of deep reinforcement learning algorithms in robotic domains by using Layer-wise Relevance Propagation

Mehran Taghian Jazi,Shotaro Miwa,Yoshihiro Mitsuka,Johannes Günther,Osmar Zaiane

A key component to the recent success of reinforcement learning is the introduction of neural networks for representation learning. Doing so allows for solving challenging problems in several domains, one of which is robotics. However, a major criticism of deep reinforcement learning (DRL) algorithms is their lack of explainability and interpretability. This problem is even exacerbated in robotics as they oftentimes cohabitate space with humans, making it imperative to be able to reason about their behaviour.

In this paper, we propose to analyze the learned representation in a robotic setting by utilizing graph neural networks. Using the graphical neural networks and Layer-wise Relevance Propagation (LRP), we represent the observations as an entity-relationship to allow us to interpret the learned policy. We evaluate our approach in two environments in MuJoCo. These two environments were delicately designed to effectively measure the value of knowledge gained by our approach to analyzing learned representations. This approach allows us to analyze not only how different parts of the observation space contribute to the decision-making process but also differentiate between policies and their differences in performance. This difference in performance also allows for reasoning about the agent's recovery from faults. These insights are key contributions to explainable deep reinforcement learning in robotic settings.

****

Learning to Take a Break: Sustainable Optimization of Long-Term User Engagement

Eden Saig,Nir Rosenfeld

Optimizing user engagement is a key goal for modern recommendation systems, but blindly pushing users towards increased consumption risks burn-out, churn, or even addictive habits. To promote digital well-being, most platforms now offer a service that periodically prompts users to take a break. These, however, must be set up manually, and so may be suboptimal for both users and the system.

In this paper, we propose a framework for optimizing long-term engagement by learning individualized breaking policies. Using Lotka-Volterra dynamics, we model users as acting based on two balancing latent states: drive, and interest---which must be conserved. We then give an efficient learning algorithm, provide theoretical guarantees, and empirically evaluate its performance on semi-synthetic data.

****

HyperQuery: A Framework for Higher Order Link Prediction

Sepideh Maleki,Josh Vekhter,Keshav Pingali

Groups with complex set intersection relations are a natural way to model a wide array of data, from the formation of social groups to the complex protein interactions which form the basis of biological life. While graphs are a natural way to represent complex networks and are well studied, typical approaches to modeling group membership using graphs are lossy. Hypergraphs are a more natural way to represent such ``higher order'' relationships, but efforts to apply machine learning techniques to hypergraph structured datasets have been limited thus far. In this paper, we address the problem of link prediction in knowledge hypergraphs as well as regular hypergraphs and develop a novel, simple, and effective optimization architecture to solve this task. Additionally, we study how integrating data from node-level labels can improve the results of our system. Our self-supervised approach achieves significant improvement over state of the art results on several hyperedge prediction and knowledge hypergraph completeion benchmarks.

****

Generative Model Based Noise Robust Training for Unsupervised Domain Adaptation

Zhongying Deng,Da Li,Junjun He,Yi-Zhe Song,Tao Xiang

Target domain pseudo-labeling has shown effectiveness in unsupervised domain adaptation (UDA). However, pseudo-labels of unlabeled target domain data are inevitably noisy due to the distribution shift between source and target domains. In this paper, we propose a generative model-based noise-robust training method (GeMo-NoRT), serving for domain shift elimination and label noise robustness simultaneously. GeMo-NoRT incorporates a distribution-based class-wise feature augmentation (D-CFA) and a generative-discriminative classifier consistency (GDC), both based on the class-wise target distributions modeled by generative models. D-CFA minimizes the domain gap by augmenting the source data with distribution-sampled target features, and trains a noise-robust discriminative classifier by using target domain knowledge from the generative models. GDC regards all the class-wise generative models as a generative classifier and enforces a consistency regularization between the generative and discriminative classifiers. It exploits an ensemble of target knowledge from all the generative models to train a noise-robust discriminative classifier, and eventually gets theoretically linked to the Ben-David domain adaptation theorem for reducing domain gap. Extensive experiments on Office-Home, PACS, and Digit-Five show that our GeMo-NoRT achieves state of the art under single-source and multi-source UDA settings.

****

Deep Learning meets Nonparametric Regression: Are Weight-Decayed DNNs Locally Adaptive?

Kaiqi Zhang,Yu-Xiang Wang

We study the theory of neural network (NN) from the lens of classical nonparametric regression problems with a focus on NN’s ability to adaptively estimate functions with heterogeneous smoothness — a property of functions in Besov or Bounded Variation (BV) classes. Existing work on this problem requires tuning the NN architecture based on the function spaces and sample sizes. We consider a “Parallel NN” variant of deep ReLU networks and show that the standard weight decay is equivalent to promoting the ■p -sparsity (0 < p < 1) of the coefficient vector of an end-to-end learned function bases, i.e., a dictionary. Using this equivalence, we further establish that by tuning only the weight decay, such Parallel NN achieves an estimation error arbitrarily close to the minimax rates for both the Besov and BV classes. Notably, it gets exponentially closer to minimax optimal as the NN gets deeper. Our research sheds new lights on why depth matters and how NNs are more powerful than kernel methods

****

Sparse Token Transformer with Attention Back Tracking

Heejun Lee,Minki Kang,Youngwan Lee,Sung Ju Hwang

Despite the success of Transformers in various applications from text, vision, and speech domains, they are yet to become standard architectures for mobile and edge device applications due to their heavy memory and computational requirements. While there exist many different approaches to reduce the complexities of the Transformers, such as the pruning of the weights/attentions/tokens, quantization, and distillation, we focus on token pruning, which reduces not only the complexity of the attention operations, but also the linear layers, which have non-negligible computational costs. However, previous token pruning approaches often remove tokens during the feed-forward stage without consideration of their impact on later layers' attentions, which has a potential risk of dropping out important tokens for the given task. To tackle this issue, we propose an attention back-tracking method that tracks the importance of each attention in a Transformer architecture from the outputs to the inputs, to preserve the tokens that have a large impact on the final predictions. We experimentally validate the effectiveness of the method on both NLP and CV benchmarks, using Transformer architectures for both domains, and the results show that the proposed attention back-tracking allows the model to better retain the full models' performance even at high sparsity rates, significantly outperforming all baselines. Qualitative analysis of the examples further shows that our method does preserve semantically meaningful tokens.

****

A Deep Conjugate Direction Method for Iteratively Solving Linear Systems

Ayano Kaneda,Osman Akar,Jingyu Chen,Victoria Alicia Trevino Kala,David Hyde,Joseph Teran

We present a novel deep learning approach to approximate the solution of large, sparse, symmetric, positive-definite linear systems of equations.These systems arise from many problems in applied science, e.g., in numerical methods for partial differential equations. Algorithms for approximating the solution to these systems are often the bottleneck in problems that require their solution, particularly for modern applications that require many millions of unknowns. Indeed, numerical linear algebra techniques have been investigated for many decades to alleviate this computational burden. Recently, data-driven techniques have also shown promise for these problems. Motivated by the conjugate gradients algorithm that iteratively selects search directions for minimizing the matrix norm of the approximation error, we design an approach that utilizes a deep neural network to accelerate convergence via data-driven improvement of the search directions.

Our method leverages a carefully chosen convolutional network to approximate the action of the inverse of the linear operator up to an arbitrary constant. We train the network using unsupervised learning with a loss function equal to the L^2 difference between an input and the system matrix times the network evaluation, where the unspecified constant in the approximate inverse is accounted for.

We demonstrate the efficacy of our approach on spatially discretized Poisson equations with millions of degrees of freedom arising in computational fluid dynamics applications.Unlike state-of-the-art learning approaches, our algorithm is capable of reducing the linear system residual to a given tolerance in a small number of iterations, independent of the problem size.Moreover, our method generalizes effectively to various systems beyond those encountered during training.

****

Smart Multi-tenant Federated Learning

Weiming Zhuang,Yonggang Wen,Shuai Zhang

Federated learning (FL) is an emerging distributed machine learning method that empowers in-situ model training on decentralized edge devices. However, multiple simultaneous training activities could overload resource-constrained devices. In this work, we propose a smart multi-tenant FL system, MuFL, to effectively coordinate and execute simultaneous training activities. We first formalize the problem of multi-tenant FL, define multi-tenant FL scenarios, and introduce a vanilla multi-tenant FL system that trains activities sequentially to form baselines. Then, we propose two approaches to optimize multi-tenant FL: 1) activity consol

idation merges training activities into one activity with a multi-task architecture; 2) after training it for rounds, activity splitting divides it into groups by employing affinities among activities such that activities within a group have better synergy. Extensive experiments demonstrate that MuFL outperforms other methods while consuming 40% less energy. We hope this work will inspire the community to further study and optimize multi-tenant FL.
****
Robust Active Distillation
Cenk Baykal,Khoa Trinh,Fotis Iliopoulos,Gaurav Menghani,Erik Vee
Distilling knowledge from a large teacher model to a lightweight one is a widely successful approach for generating compact, powerful models in the semi-supervised learning setting where a limited amount of labeled data is available. In large-scale applications, however, the teacher tends to provide a large number of incorrect soft-labels that impairs student performance. The sheer size of the teacher additionally constrains the number of soft-labels that can be queried due to prohibitive computational and/or financial costs. The difficulty in achieving simultaneous \emph{efficiency} (i.e., minimizing soft-label queries) and \emph{robustness} (i.e., avoiding student inaccuracies due to incorrect labels) hurts the widespread application of knowledge distillation to many modern tasks. In this paper, we present a parameter-free approach with provable guarantees to query the soft-labels of points that are simultaneously informative and correctly labeled by the teacher. At the core of our work lies a game-theoretic formulation that explicitly considers the inherent trade-off between the informativeness and correctness of input instances. We establish bounds on the expected performance of our approach that hold even in worst-case distillation instances. We present empirical evaluations on popular benchmarks that demonstrate the improved distillation performance enabled by our work relative to that of state-of-the-art active learning and active distillation methods.
****
Controllable Image Generation via Collage Representations
Arantxa Casanova,Marlene Careil,Adriana Romero-Soriano,Christopher Pal,Jakob Verbeek,Michal Drozdzal
Recent advances in conditional generative image models have enabled impressive results. On the one hand, text-based conditional models have achieved remarkable generation quality, by leveraging large-scale datasets of image-text pairs. To enable fine-grained controllability, however, text-based models require long prompts, whose details may be ignored by the model. On the other hand, layout-based conditional models have also witnessed significant advances. These models rely on bounding boxes or segmentation maps for precise spatial conditioning in combination with coarse semantic labels. The semantic labels, however, cannot be used to express detailed appearance characteristics. In this paper, we approach fine-grained scene controllability through image collages which allow a rich visual description of the desired scene as well as the appearance and location of the objects therein, without the need of class nor attribute labels. We introduce "mixing and matching scenes" (M&Ms), an approach that consists of an adversarially trained generative image model which is conditioned on appearance features and spatial positions of the different elements in a collage, and integrates these into a coherent image. We train our model on the OpenImages (OI) dataset and evaluate it on collages derived from OI and MS-COCO datasets. Our experiments on the OI dataset show that M&Ms outperforms baselines in terms of fine-grained scene controllability while being very competitive in terms of image quality and sample diversity. On the MS-COCO dataset, we highlight the generalization ability of our model by outperforming DALL-E in terms of the zero-shot FID metric, despite using two magnitudes fewer parameters and data. Collage based generative models have the potential to advance content creation in an efficient and effective way as they are intuitive to use and yield high quality generations.
****
Robust Multi-Agent Reinforcement Learning with State Uncertainties
Sihong He,Songyang Han,Sanbao Su,Shuo Han,Shaofeng Zou,Fei Miao
In real-world multi-agent reinforcement learning (MARL) applications, agents may

not have perfect state information (e.g., due to inaccurate measurement or malicious attacks), which challenges the robustness of agents' policies. Though robustness is getting important in MARL deployment, little prior work has studied state uncertainties in MARL, neither in problem formulation nor algorithm design. Motivated by this robustness issue, we study the problem of MARL with state uncertainty in this work. We provide the first attempt to the theoretical and empirical analysis of this challenging problem. We first model the problem as a Markov Game with state perturbation adversaries (MG-SPA), and introduce Robust Equilibrium as the solution concept. We conduct fundamental analysis regarding MG-SPA and give conditions under which such an equilibrium exists. Then we propose a robust multi-agent Q-learning (RMAQ) algorithm to find such an equilibrium, with convergence guarantees. To handle high-dimensional state-action space, we design a robust multi-agent actor-critic (RMAAC) algorithm based on an analytical expression of the policy gradient derived in the paper. Our experiments show that the proposed RMAQ algorithm converges to the optimal value function; our RMAAC algorithm outperforms several MARL methods that do not consider the state uncertainty in several multi-agent environments.

****

Tiny Adapters for Vision Transformers

Imad Eddine MAROUF,Enzo Tartaglione,Stéphane Lathuilière

Vision Transformers (ViTs) have become one of the dominant architectures in computer vision and pretrained ViT models are commonly adapted to new tasks via fine-tuning of its parameters. Recent works in NLP proposed a variety of parameter-efficient transfer learning methods such as adapters to avoid the prohibitive storage cost of fine-tuning.

In this work, we start from the observation that adapters perform poorly when the dimension of adapters is small and we propose a training algorithm that addresses this issue. We start from large adapters which can be trained easily and iteratively reduce the size of every adapter. We introduce a scoring function that can compare neuron importance across layers and consequently allow automatic estimation of the hidden dimension of every adapter. Our method outperforms existing approaches in terms of the trade-off between accuracy and trained parameters across domain adaptation benchmarks. We will release our code publicly upon acceptance.

****

Accelerating Inverse Reinforcement Learning with Expert Bootstrapping

David Wu,Sanjiban Choudhury

Existing inverse reinforcement learning methods (e.g. MaxEntIRL, f-IRL) search over candidate reward functions and solve a reinforcement learning problem in the inner loop. This creates a rather strange inversion where a harder problem, reinforcement learning, is in the inner loop of a presumably easier problem, imitation learning. In this work, we show that better utilization of expert demonstrations can reduce the need for hard exploration in the inner RL loop, hence accelerating learning. Specifically, we propose two simple recipes: (1) placing expert transitions into the replay buffer of the inner RL algorithm (e.g. Soft-Actor Critic) which directly informs the learner about high reward states instead of forcing the learner to discover them through extensive exploration, and (2) using expert actions in Q value bootstrapping in order to improve the target Q value estimates and more accurately describe high value expert states. Our methods show significant gains over a MaxEntIRL baseline on the benchmark MuJoCo suite of tasks, speeding up recovery to 70\% of deterministic expert performance by 2.13x on HalfCheetah-v2, 2.6x on Ant-v2, 18x on Hopper-v2, and 3.36x on Walker2d-v2.

****

Kernel Neural Optimal Transport

Alexander Korotin,Daniil Selikhanovych,Evgeny Burnaev

We study the Neural Optimal Transport (NOT) algorithm which uses the general optimal transport formulation and learns stochastic transport plans. We show that NOT with the weak quadratic cost may learn fake plans which are not optimal. To resolve this issue, we introduce kernel weak quadratic costs. We show that they p

rovide improved theoretical guarantees and practical performance. We test NOT with kernel costs on the unpaired image-to-image translation task.
****

Neural Optimal Transport
Alexander Korotin,Daniil Selikhanovych,Evgeny Burnaev
We present a novel neural-networks-based algorithm to compute optimal transport maps and plans for strong and weak transport costs. To justify the usage of neural networks, we prove that they are universal approximators of transport plans between probability distributions. We evaluate the performance of our optimal transport algorithm on toy examples and on the unpaired image-to-image translation.
****

SeaFormer: Squeeze-enhanced Axial Transformer for Mobile Semantic Segmentation
Qiang Wan,Zilong Huang,Jiachen Lu,Gang YU,Li Zhang
Since the introduction of Vision Transformers, the landscape of many computer vision tasks (e.g., semantic segmentation), which has been overwhelmingly dominated by CNNs, recently has significantly revolutionized. However, the computational cost and memory requirement render these methods unsuitable on the mobile device, especially for the high resolution per-pixel semantic segmentation task. In this paper, we introduce a new method squeeze-enhanced Axial Transformer (SeaFormer) for mobile semantic segmentation. Specifically, we design a generic attention block characterized by the formulation of squeeze Axial and spatial enhancement. It can be further used to create a family of backbone architectures with superior cost-effectiveness. Coupled with a light segmentation head, we demonstrate state-of-the-art results on the ADE20K, Pascal Context and COCO-stuff datasets. Critically, we beat both the mobile-friendly rivals and Transformer-based counterparts with better performance and lower latency without bells and whistles. Beyond semantic segmentation, we further apply the proposed SeaFormer architecture to image classification problem, demonstrating the potentials of serving as a versatile mobile-friendly backbone.
****

Joint Edge-Model Sparse Learning is Provably Efficient for Graph Neural Networks
Shuai Zhang,Meng Wang,Pin-Yu Chen,Sijia Liu,Songtao Lu,Miao Liu
Due to the significant computational challenge of training large-scale graph neural networks (GNNs), various sparse learning techniques have been exploited to reduce memory and storage costs. Examples include graph sparsification that samples a subgraph to reduce the amount of data aggregation and model sparsification that prunes the neural network to reduce the number of trainable weights. Despite the empirical successes in reducing the training cost while maintaining the test accuracy, the theoretical generalization analysis of sparse learning for GNNs remains elusive. To the best of our knowledge, this paper provides the first theoretical characterization of joint edge-model sparse learning from the perspective of sample complexity and convergence rate in achieving zero generalization error. It proves analytically that both sampling important nodes and pruning neurons with lowest-magnitude can reduce the sample complexity and improve convergence without compromising the test accuracy. Although the analysis is centered on two-layer GNNs with structural constraints on data, the insights are applicable to more general setups and justified by both synthetic and practical citation datasets.
****

Harnessing spectral representations for subgraph alignment
Marco Pegoraro,Riccardo Marin,Arianna Rampini,Simone Melzi,Luca Cosmo,Emanuele Rodolà
With the rise and advent of graph learning techniques, graph data has become ubiquitous in the machine learning field. However, while several efforts have been devoted to the design of new convolutional architectures, pooling or positional encoding schemes, relatively little focus has been spent on modeling pairwise problems such as signal transfer, graph isomorphism and subgraph correspondence tasks. With this paper, we anticipate the need for a convenient framework to deal with problems that revolve around the notion of a map among graphs, and focus in particular on the challenging subgraph alignment scenario. We claim that, first

and foremost, the representation of a map plays a central role in how these problems should be modeled -- be it a map inference problem or a simpler signal transport task. Taking the hint from recent work in geometry processing, we propose the adoption of a spectral representation for maps that is compact, easy to compute, permutation-equivariant, easy to plug into learning pipelines, and especially effective for a wide range of situations, most notably when dealing with subgraph alignment problems. We further report for the first time a surprising phenomenon where the partiality arising in subgraph alignment is manifested in the structure of the map coefficients, even in the absence of exact isomorphism, and which is consistently observed over different families of graphs.

****

MotifExplainer: a Motif-based Graph Neural Network Explainer

Zhaoning Yu,Hongyang Gao

We consider the explanation problem of Graph Neural Networks (GNNs). Most existing GNN explanation methods identify the most important edges or nodes but fail to consider substructures, which are more important for graph data. One method considering subgraphs tries to search all possible subgraphs and identifies the most significant ones. However, the subgraphs identified may not be recurrent or statistically important for interpretation. This work proposes a novel method, named MotifExplainer, to explain GNNs by identifying important motifs, which are recurrent and statistically significant patterns in graphs. Our proposed motif-based methods can provide better human-understandable explanations than methods based on nodes, edges, and regular subgraphs. Given an instance graph and a pre-trained GNN model, our method first extracts motifs in the graph using domain-specific motif extraction rules. Then, a motif embedding is encoded by feeding motifs into the pre-trained GNN. Finally, we employ an attention-based method to identify the most influential motifs as explanations for the prediction results. The empirical studies on both synthetic and real-world datasets demonstrate the effectiveness of our method.

****

Receding Neuron Importances for Structured Pruning

Mihai Suteu,Yike Guo

Structured pruning efficiently compresses networks by identifying and removing unimportant neurons. While this can be elegantly achieved by applying sparsity-inducing regularisation on BatchNorm parameters, an L1 penalty would shrink all scaling factors rather than just those of superfluous neurons. To tackle this issue, we introduce a simple BatchNorm variation with bounded scaling parameters, based on which we design a novel regularisation term that suppresses only neurons with low importance. Under our method, the weights of unnecessary neurons effectively recede, producing a polarised bimodal distribution of importances. We show that neural networks trained this way can be pruned to a larger extent and with less deterioration. We one-shot prune VGG and ResNet architectures at different ratios on CIFAR and ImagenNet datasets. In the case of VGG-style networks, our method significantly outperforms existing approaches particularly under severe pruning.

****

Learning Sparse and Low-Rank Priors for Image Recovery via Iterative Reweighted Least Squares Minimization

Stamatios Lefkimmiatis,Iaroslav Sergeevich Koshelev

In this work we introduce a novel optimization algorithm for image recovery under learned sparse and low-rank constraints, which are parameterized with weighted extensions of the ℓ_p^p-vector and $\mathcal{S}_p^p$ Schatten-matrix quasi-norms for $0\!<p\!\le1$, respectively. Our proposed algorithm generalizes the Iteratively Reweighted Least Squares (IRLS) method, used for signal recovery under ℓ_1 and nuclear-norm constrained minimization. Further, we interpret our overall minimization approach as a recurrent network that we then employ to deal with inverse low-level computer vision problems. Thanks to the convergence guarantees that our IRLS strategy offers, we are able to train the derived reconstruction networks using a memory-efficient implicit back-propagation scheme, which does not pose any restrictions on their effective depth. To assess our networks'

performance, we compare them against other existing reconstruction methods on several inverse problems, namely image deblurring, super-resolution, demosaicking and sparse recovery. Our reconstruction results are shown to be very competitive and in many cases outperform those of existing unrolled networks, whose number of parameters is orders of magnitude higher than that of our learned models.

****

Unifying Diffusion Models' Latent Space, with Applications to CycleDiffusion and Guidance

Chen Henry Wu,Fernando De la Torre

Diffusion models have achieved unprecedented performance in generative modeling. The commonly-adopted formulation of the latent code of diffusion models is a sequence of gradually denoised samples, as opposed to the simpler (e.g., Gaussian) latent space of GANs, VAEs, and normalizing flows. This paper provides an alternative, Gaussian formulation of the latent space of various diffusion models, as well as an invertible DPM-Encoder that maps images into the latent space. While our formulation is purely based on the definition of diffusion models, we demonstrate several intriguing consequences. (1) Empirically, we observe that a common latent space emerges from two diffusion models trained independently on related domains. In light of this finding, we propose CycleDiffusion, which uses DPM-Encoder for unpaired image-to-image translation. Furthermore, applying CycleDiffusion to text-to-image diffusion models, we show that large-scale text-to-image diffusion models can be used as zero-shot image-to-image editors. (2) One can guide pre-trained diffusion models and GANs by controlling the latent codes in a unified, plug-and-play formulation based on energy-based models. Using the CLIP model and a face recognition model as guidance, we demonstrate that diffusion models have better coverage of low-density sub-populations and individuals than GANs.

****

Spherical Sliced-Wasserstein

Clément Bonet,Paul Berg,Nicolas Courty,François Septier,Lucas Drumetz,Minh Tan Pham

Many variants of the Wasserstein distance have been introduced to reduce its original computational burden. In particular the Sliced-Wasserstein distance (SW), which leverages one-dimensional projections for which a closed-form solution of the Wasserstein distance is available, has received a lot of interest. Yet, it is restricted to data living in Euclidean spaces, while the Wasserstein distance has been studied and used recently on manifolds. We focus more specifically on the sphere, for which we define a novel SW discrepancy, which we call spherical Sliced-Wasserstein, making a first step towards defining SW discrepancies on manifolds. Our construction is notably based on closed-form solutions of the Wasserstein distance on the circle, together with a new spherical Radon transform. Along with efficient algorithms and the corresponding implementations, we illustrate its properties in several machine learning use cases where spherical representations of data are at stake: sampling on the sphere, density estimation on real eath data or hyperspherical auto-encoders.

****

Intepreting & Improving Pretrained Language Models: A Probabilistic Conceptual Approach

Hengyi Wang,Zhiqing Hong,Desheng Zhang,Hao Wang

Pretrained Language Models (PLMs) such as BERT and its variants have achieved remarkable success in natural language processing. To date, the interpretability of PLMs has primarily relied on the attention weights in their self-attention layers. However, these attention weights only provide word-level interpretations, failing to capture higher-level structures, and are therefore lacking in readability and intuitiveness. In this paper, we propose a hierarchical Bayesian deep learning model, dubbed continuous latent Dirichlet allocation (CLDA), to go beyond word-level interpretations and provide concept-level interpretations. Our CLDA is compatible with any attention-based PLMs and can work as either (1) an interpreter which interprets model predictions at the concept level without any performance sacrifice or (2) a regulator which is jointly trained with PLMs during fin

etuning to further improve performance. Experimental results on various benchmark datasets show that our approach can successfully provide conceptual interpretation and performance improvement for state-of-the-art PLMs.
****

Neural Optimal Transport with General Cost Functionals

Arip Asadulaev,Alexander Korotin,Vage Egiazarian,Evgeny Burnaev

We present a novel neural-networks-based algorithm to compute optimal transport (OT) plans and maps for general cost functionals. The algorithm is based on a saddle point reformulation of the OT problem and generalizes prior OT methods for weak and strong cost functionals. As an application, we construct a functional to map data distributions with preserving the class-wise structure of data.
****

Does Dataset Lottery Ticket Hypothesis Exist?

Zhiqiang Shen,Eric Xing

Tuning hyperparameters and exploring the suitable training schemes for the self-supervised models are usually expensive and resource-consuming, especially on large-scale datasets like ImageNet-1K. Critically, this means only a few establishments (e.g., Google, Meta, etc.) have the ability to afford the heavy experiments on this task, which seriously hinders more engagement and better development of this area. An ideal situation is that there exists a subset from the full large-scale dataset, the subset can correctly reflect the performance distinction when performing different training frameworks, hyper-parameters, etc. This new training manner will substantially decrease resource requirements and improve the computational performance of ablations without compromising accuracy on the full dataset. We formulate this interesting problem as the dataset lottery ticket hypothesis and the target subsets as the winning tickets.

In this work, we analyze this problem through finding out partial empirical data on the class dimension that has a consistent {\em Empirical Risk Trend} as the full observed dataset. We also examine multiple solutions, including (i) a uniform selection scheme that has been widely used in literature; (ii) subsets by involving prior knowledge, for instance, using the sorted per-class performance of the strong supervised model to identify the desired subset, WordNet Tree on hierarchical semantic classes, etc., for generating the target winning tickets.

We verify this hypothesis on the self-supervised learning task across a variety of recent mainstream methods, such as MAE, DINO, MoCo-V1/V2, etc., with different backbones like ResNet and Vision Transformers. The supervised classification task is also examined as an extension. We conduct extensive experiments for training more than 2K self-supervised models on the large-scale ImageNet-1K and its subsets by 1.5M GPU hours, to scrupulously deliver our discoveries and demonstrate our conclusions. According to our experimental results, the winning tickets (subsets) that we find behave consistently to the original dataset, which generally can benefit many experimental studies and ablations, saving 10x of training time and resources for the hyperparameter tuning and other ablation studies.
****

Random Weight Factorization improves the training of Continuous Neural Representations

Sifan Wang,Hanwen Wang,Jacob H Seidman,Paris Perdikaris

Continuous neural representations have recently emerged as a powerful and flexible alternative to classical discretized representations of signals. However, training them to capture fine details in multi-scale signals is difficult and computationally expensive. Here we propose random weight factorization as a simple drop-in replacement for parameterizing and initializing conventional linear layers in coordinate-based multi-layer perceptrons (MLPs) that significantly accelerates and improves their training. We show how this factorization alters the underlying loss landscape and effectively enables each neuron in the network to learn using its own self-adaptive learning rate. This not only helps with mitigating spectral bias, but also allows networks to quickly recover from poor initializations and reach better local minima. We demonstrate how random weight factorizatio n can be leveraged to improve the training of neural representations on a variety of tasks, including image regression, shape representation, computed tomography, inverse rendering, solving partial differential equations, and learning operators between function spaces.

****

InPL: Pseudo-labeling the Inliers First for Imbalanced Semi-supervised Learning

Zhuoran Yu,Yin Li,Yong Jae Lee

Recent state-of-the-art methods in imbalanced semi-supervised learning (SSL) rely on confidence-based pseudo-labeling with consistency regularization. To obtain high-quality pseudo-labels, a high confidence threshold is typically adopted. However, it has been shown that softmax-based confidence scores in deep networks can be arbitrarily high for samples far from the training data, and thus, the pseudo-labels for even high-confidence unlabeled samples may still be unreliable. In this work, we present a new perspective of pseudo-labeling for imbalanced SSL. Without relying on model confidence, we propose to measure whether an unlabeled sample is likely to be "in-distribution''; i.e., close to the current training data. To decide whether an unlabeled sample is "in-distribution'' or "out-of-distribution'', we adopt the energy score from out-of-distribution detection literature. As training progresses and more unlabeled samples become in-distribution and contribute to training, the combined labeled and pseudo-labeled data can better approximate the true class distribution to improve the model. Experiments demonstrate that our energy-based pseudo-labeling method, InPL, albeit conceptually simple, significantly outperforms confidence-based methods on imbalanced SSL benchmarks. For example, it produces a 4-6% absolute accuracy improvement on CIFAR10-LT when the imbalance ratio is higher than 50. When combined with state-of-the-art long-tailed SSL methods, further improvements are attained. In particular, in one of the most challenging scenarios, InPL achieves a 6.9% accuracy improvement over the best competitor.

****

Mixed-Precision Inference Quantization: Problem Resetting, Mapping math concept and Branch\&bound methods

Daning Cheng

Based on the model's resilience to computational noise, model quantization is important for compressing models and improving computing speed. Existing quantization techniques rely heavily on experience and "fine-tuning" skills.   This study shows that in inference process, the mixed-precision quantization is a NP-hard problem and we designed a set of method to map mathematical method into pratical method. And our problem setting can be solved by branch and bound method with less computing resouces. We also show that how to set the quantization parameters in theorical method.

****

Latent Offline Distributional Actor-Critic

Félicien Hêche,BARAKAT Oussama,Thibaut P Desmettre,Tania Marx,Stephan Robert-Nicoud

Offline reinforcement learning (RL) has emerged as a promising paradigm for real world applications, since it aims to train policies directly from datasets of past interactions with the environment. The past few years, algorithms have been introduced to learn policies from high-dimensional observational states in an offline settings. The general idea of these methods is to encode the environment into a smaller latent space and train policies on the top of this smaller representation. In this paper, we extend this general method to stochastic environments (i.e. where the reward function is stochastic) and considering a risk measure instead of the classical expected return. First, we show that under some assumptions it is equivalent to minimize a risk measure in the latent space and in the natural space. Based on this result, we present Latent Offline Distributional Actor-Critic (LODAC), an algorithm which is able to train policies in high-dimensional stochastic and offline settings to minimize a given risk measure. Empirically, we show that using LODAC to minimize Conditional Value-at-Risk (CVaR), outperforms previous methods in term of CVaR and return on stochastic environments.

****

Mixed-Precision Inference Quantization: Radically Towards Faster inference speed, Lower Storage requirement, and Lower Loss

Daning Cheng

Model quantization is important for compressing models and improving computing speed. However, current researchers think that the loss function of quantizated model is usually higher than full precision model. This study provides a methodology for acquiring a mixed-precise quantization model with a lower loss than the full precision model. In addition, the analysis demonstrates that, throughout the inference process, the loss function is mostly affected by the noise of the layer inputs. In particular, we will demonstrate that neural networks with massive identity mappings are resistant to the quantization method. It is also difficult to improve the performance of these networks using quantization.

****

Leveraging Double Descent for Scientific Data Analysis: Face-Based Social Behavior as a Case Study

Christine H Lind,Angela J. Yu

Scientific data analysis often involves making use of a large number of correlated predictor variables to predict multiple response variables. Understanding how the predictor and response variables relate to one another, especially in the presence of relatively scarce data, is a common and challenging problem. Here, we leverage the recently popular concept of ``double descent'' to develop a particular treatment of the problem, including a set of key theoretical results. We also apply the proposed method to a novel experimental dataset consisting of human ratings of social traits and social decision making tendencies based on the facial features of strangers, and resolve a scientific debate regarding the existence of a ``beauty premium'' or ``attractiveness halo,'' which refers to a (presumed) advantage attractive people enjoy in social situations. We demonstrate that more attractive faces indeed enjoy a social advantage, but this is indirectly due to the facial features that contribute to both perceived attractiveness and trustworthiness, and that the component of attractiveness perception due to facial features (unrelated to trustworthiness) actually elicit a ``beauty penalty'', which has also been reported in the literature. Conversely, the facial features that contribute to trustworthiness and not to attractiveness still contribute positively to pro-social trait perception and decision making. Thus, what was previously thought to be an ``attractiveness halo'' is actually a ``trustworthiness halo'' plus a ``beauty penalty.'' Moreover, we see that the facial features that contribute to the ``trustworthiness halo' primarily have to do with how smiley a face is, while the facial features that contribute to attractiveness but actually acts as a ``beauty penalty'' is anti-correlated with age. In other words, youthfulness and smiley-ness both contribute to attractiveness, but only smiley-ness positively contributes to pro-social perception and decision making, while youthfulness actually negatively contribute to them. A further interesting wrinkle is that youthfulness as a whole does not negatively contribute to social traits/decision-making, only the component of youthfulness contributing to attractiveness does.

****

Maximizing Communication Efficiency for Large-scale Training via 0/1 Adam

Yucheng Lu,Conglong Li,Minjia Zhang,Christopher De Sa,Yuxiong He

1-bit gradient compression and local steps are two representative techniques that enable drastic communication reduction in distributed SGD. Their benefits, however, remain an open question on Adam-based large model pre-training (e.g. BERT and GPT). In this paper, we demonstrate the non-linearity in Adam causes slow convergence even when 1-bit compression or local steps are individually applied. To alleviate this limitation, we propose \textbf{0/1 Adam} that linearizes each Adam step via approximating its optimizer states using their stale estimates and linear correlation. \textbf{0/1 Adam} performs an Adam-like step to preserve the adaptivity, while its linearity allows utilizing 1-bit compression and local steps simultaneously for wall-clock time speed up. We provide convergence guarantee for \textbf{0/1 Adam} on smooth non-convex objectives. On various large-scale benchmarks such as BERT-Base, BERT-Large, GPT-2 pre-training and ImageNet, we demonstrate on up to 128 GPUs that \textbf{0/1 Adam} is able to reduce up to 87\% of data volume, 54\% of communication rounds, and achieve up to 2$\times$ higher training throughput and end-to-end training time reduction compared to the state-of-the-art baseline 1-bit Adam; while enjoying the same statistical convergence speed and end task model accuracy on GLUE dataset and ImageNet validation set.

****

Truthful Self-Play

Shohei Ohsawa

We present a general framework for evolutionary learning to emergent unbiased state representation without any supervision. Evolutionary frameworks such as self-play converge to bad local optima in case of multi-agent reinforcement learning in non-cooperative partially observable environments with communication due to information asymmetry. Our proposed framework is a simple modification of self-play inspired by mechanism design, also known as {\em reverse game theory}, to elicit truthful signals and make the agents cooperative. The key idea is to add imaginary rewards using the peer prediction method, i.e., a mechanism for evaluating the validity of information exchanged between agents in a decentralized environment. Numerical experiments with predator prey, traffic junction and StarCraft tasks demonstrate that the state-of-the-art performance of our framework.

****

Iterative α-(de)Blending: Learning a Deterministic Mapping Between Arbitrary Densities

Eric Heitz,Laurent Belcour,Thomas Chambon

We present a learning method that produces a mapping between arbitrary densities, such that random samples of a density can be mapped to random samples of another. In practice, our method is similar to deterministic diffusion processes where samples of the target density are blended with Gaussian noise. The originality of our approach is that, in contrast to several recent works, we do not rely on Langevin dynamics or score-matching concepts. We propose a simpler take on the topic, which is based solely on Bayes' theorem. By studying blended samples and their posteriors, we show that iteratively blending and deblending samples produces random paths between arbitrary densities. We prove that, for finite-variance densities, these paths converge towards a deterministic mapping that can be learnt with a neural network trained to deblend samples. Our method can thus be seen as a generalization of deterministic denoising diffusion where, instead of learning to denoise Gaussian noise, we learn to deblend arbitrary data.

****

Strategic Classification with Graph Neural Networks

Itay Eilat,Ben Finkelshtein,Chaim Baskin,Nir Rosenfeld

Strategic classification studies learning in settings where users can modify their features to obtain favorable predictions. Most current works focus on simple classifiers that trigger independent user responses. Here we examine the implications of learning with more elaborate models that break the independence assumption. Motivated by the idea that applications of strategic classification are often social in nature, we focus on graph neural networks, which make use of social relations between users to improve predictions. Using a graph for learning introduces inter-user dependencies in prediction; our key point is that strategic users can exploit these to promote their goals. As we show through analysis and simulation, this can work either against the system---or for it. Based on this, we propose a differentiable framework for strategically-robust learning of graph-based classifiers. Experiments on several real networked datasets demonstrate the utility of our approach.

****

Continual Transformers: Redundancy-Free Attention for Online Inference

Lukas Hedegaard,Arian Bakhtiarnia,Alexandros Iosifidis

Transformers in their common form are inherently limited to operate on whole token sequences rather than on one token at a time. Consequently, their use during online inference on time-series data entails considerable redundancy due to the overlap in successive token sequences. In this work, we propose novel formulations of the Scaled Dot-Product Attention, which enable Transformers to perform efficient online token-by-token inference on a continual input stream. Importantly, our modifications are purely to the order of computations, while the outputs and learned weights are identical to those of the original Transformer Encoder. We validate our Continual Transformer Encoder with experiments on the THUMOS14, TVSeries and GTZAN datasets with remarkable results: Our Continual one- and two-block architectures reduce the floating point operations per prediction by up to 63x and 2.6x, respectively, while retaining predictive performance.

****

FedPSE: Personalized Sparsification with Element-wise Aggregation for Federated Learning

Longfei Zheng,Yingting Liu,Xiaolong Xu,Chaochao Chen,Weipeng Sun,Xiaolong Hu,Lei Wang,Li Wang

Federated learning (FL) is a popular distributed machine learning framework in which clients aggregate models' parameters instead of sharing their individual data. In FL, clients communicate with the server under limited network bandwidth frequently, which arises the communication challenge. To resolve this challenge, multiple compression methods have been proposed to reduce the transmitted parameters. However, these techniques show that the federated performance degrades significantly with Non-IID (non-identically independently distributed) datasets. To address this issue, we propose an effective method, called FedPSE, which solves the efficiency challenge of FL with heterogeneous data. FedPSE compresses the local updates on clients using Top-K sparsification and aggregates these updates on the server by element-wise average. Then clients download the personalized sparse updates from the server to update their individual local models. We then theoretically analyze the convergence of FedPSE under the non-convex setting. Moreover, extensive experiments on four benchmark tasks demonstrate that our FedPSE outperforms the state-of-the-art methods on Non-IID datasets in terms of both efficiency and accuracy.

****

Towards Online Real-Time Memory-based Video Inpainting Transformers

Guillaume Thiry,Hao Tang,Radu Timofte,Luc Van Gool

Video inpainting tasks have seen significant improvements in the past years with the rise of deep neural networks and, in particular, vision transformers. Although these models show promising reconstruction quality and temporal consistency, they are still unsuitable for live videos, one of the last steps to make them completely convincing and usable. The main limitations are that these state-of-the-art models inpaint using the whole video (offline processing) and show an insufficient frame rate. In our approach, we propose a framework to adapt existing inpainting transformers to these constraints by memorizing and refining redundant computations while maintaining a decent inpainting quality. Using this framework with some of the most recent inpainting models, we show great online results with a consistent throughput above 20 frames per second. Code and pretrained models will be made available upon acceptance.

****

Edge-Varying Fourier Graph Network for Multivariate Time Series Forecasting

Kun Yi,Qi Zhang

The key problem in multivariate time series (MTS) analysis and forecasting aims to disclose the underlying couplings between variables that drive the co-movements. Considerable recent successful MTS methods are built with graph neural networks (GNNs) due to their essential capacity for relational modeling. However, previous work often used a static graph structure of time-series variables for modeling MTS failing to capture their ever-changing correlations over time. To this end, a fully-connected supra-graph connecting any two variables at any two timestamps is adaptively learned to capture the high-resolution variable dependencies via an efficient graph convolutional network. Specifically, we construct the Edge-Varying Fourier Graph Networks (EV-FGN) equipped with Fourier Graph Shift Operator (FGSO) which efficiently performs graph convolution in the frequency domain. As a result, a high-efficiency scale-free parameter learning scheme is derive d for MTS analysis and forecasting according to the convolution theorem. Extensive experiments show that EV-FGN outperforms state-of-the-art methods on seven real-world MTS datasets.

****

Learning Symbolic Models for Graph-structured Physical Mechanism
Hongzhi Shi,Jingtao Ding,Yufan Cao,quanming yao,Li Liu,Yong Li

Graph-structured physical mechanisms are ubiquitous in real-world scenarios, thus revealing underneath formulas is of great importance for scientific discovery. However, classical symbolic regression methods fail on this task since they can only handle input-output pairs that are not graph-structured. In this paper, we propose a new approach that generalizes symbolic regression to graph-structured physical mechanisms. The essence of our method is to model the formula skeleton with a message-passing flow, which helps transform the discovery of the skeleton into the search for the message-passing flow. Such a transformation guarantees that we are able to search a message-passing flow, which is efficient and Pareto-optimal in terms of both accuracy and simplicity. Subsequently, the underneath formulas can be identified by interpreting component functions of the searched message-passing flow, reusing classical symbolic regression methods. We conduct extensive experiments on datasets from different physical domains, including mechanics, electricity, and thermology, and on real-world datasets of pedestrian dynamics without ground-truth formulas. The experimental results not only verify the rationale of our design but also demonstrate that the proposed method can automatically learn precise and interpretable formulas for graph-structured physical mechanisms.

****

Unleashing Mask: Explore the Intrinsic Out-of-distribution Detection Capability
Jianing Zhu,Hengzhuang Li,Jiangchao Yao,Tongliang Liu,Jianliang Xu,Bo Han

Out-of-distribution (OOD) detection is an important aspect for safely deploying machine learning models in real-world applications. Previous approaches either design better scoring functions or utilize the knowledge from unknown outliers to equip the well-trained models with the ability of OOD detection. However, few of them explore to excavate the intrinsic OOD detection capability of a given model. In this work, we discover the existence of an intermediate stage of a model trained on in-distribution data having higher OOD detection performance than that of its final stage across different settings, and further identify the critical attribution to be learning with atypical samples. Based on such empirical insights, we propose a new method, Unleashing Mask (UM), to reveal the once-covered detection capability of a given model. To be specific, we utilize the mask to figure out the memorized atypical samples and fine-tune the model to forget them. Extensive experiments have been conducted to characterize and verify the effectiveness of our method.

****

Dirichlet-based Uncertainty Calibration for Active Domain Adaptation
Mixue Xie,Shuang Li,Rui Zhang,Chi Harold Liu

Active domain adaptation (DA) aims to maximally boost the model adaptation on a new target domain by actively selecting limited target data to annotate, whereas traditional active learning methods may be less effective since they do not consider the domain shift issue. Despite active DA methods address this by further proposing targetness to measure the representativeness of target domain characteristics, their predictive uncertainty is usually based on the prediction of deterministic models, which can easily be miscalibrated on data with distribution shift. Considering this, we propose a Dirichlet-based Uncertainty Calibration (DUC) approach for active DA, which simultaneously achieves the mitigation of miscalibration and the selection of informative target samples. Specifically, we place a Dirichlet prior on the prediction and interpret the prediction as a distribution on the probability simplex, rather than a point estimate like deterministic models. This manner enables us to consider all possible predictions, mitigating the miscalibration of unilateral prediction. Then a two-round selection strategy based on different uncertainty origins is designed to select target samples that are both representative of target domain and conducive to discriminability. Ex tensive experiments on cross-domain image classification and semantic segmentation validate the superiority of DUC.

****

Accurate Image Restoration with Attention Retractable Transformer

Jiale Zhang,Yulun Zhang,Jinjin Gu,Yongbing Zhang,Linghe Kong,Xin Yuan

Recently, Transformer-based image restoration networks have achieved promising improvements over convolutional neural networks due to parameter-independent global interactions. To lower computational cost, existing works generally limit self-attention computation within non-overlapping windows. However, each group of tokens are always from a dense area of the image. This is considered as a dense attention strategy since the interactions of tokens are restrained in dense regions. Obviously, this strategy could result in restricted receptive fields. To address this issue, we propose \textbf{A}ttention \textbf{R}etractable \textbf{T}ransformer (ART) for image restoration, which presents both dense and sparse attention modules in the network. The sparse attention module allows tokens from sparse areas to interact and thus provides a wider receptive field. Furthermore, the alternating application of dense and sparse attention modules greatly enhances representation ability of Transformer while providing retractable attention on the input image.We conduct extensive experiments on image super-resolution, denoising, and JPEG compression artifact reduction tasks. Experimental results validate that our proposed ART outperforms state-of-the-art methods on various benchmark datasets both quantitatively and visually. We also provide code and models at~\url{https://github.com/gladzhang/ART}.

****

Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models

Xingyu Xie,Pan Zhou,Huan Li,Zhouchen Lin,Shuicheng YAN

Adaptive gradient algorithms combine the moving average idea with heavy ball acceleration to estimate accurate first- and second-order moments of the gradient for accelerating convergence. However, Nesterov acceleration which converges faster than heavy ball acceleration in theory and also in many empirical cases, is much less investigated under the adaptive gradient setting. In this work, we propose the ADAptive Nesterov momentum algorithm, Adan for short, to speed up the training of deep neural networks effectively. Adan first reformulates the vanilla Nesterov acceleration to develop a new Nesterov momentum estimation (NME) method, which avoids the extra computation and memory overhead of computing gradient at the extrapolation point. Then Adan adopts NME to estimate the first- and second-order moments of the gradient in adaptive gradient algorithms for convergence acceleration. Besides, we prove that Adan finds an ϵ-approximate first-order stationary point within $O(\epsilon^{-3.5})$ stochastic gradient complexity on the non-convex stochastic problems (e.g., deep learning problems), matching the best-known lower bound. Extensive experimental results show that Adan surpasses the corresponding SoTA optimizers on vision, language, and RL tasks and sets new SoTAs for many popular networks and frameworks, e.g., ResNet, ConvNext, ViT, Swin, MAE, LSTM, Transformer-XL, and BERT. More surprisingly, Adan can use half of the training cost (epochs) of SoTA optimizers to achieve higher or comparable performance on ViT, ResNet, MAE, etc., and also shows great tolerance to a large range of minibatch size, e.g., from 1k to 32k. We hope Adan can contribute to developing deep learning by reducing training costs and relieving the engineering burden of trying different optimizers on various architectures.

****

Efficient Trojan Injection: 90% Attack Success Rate Using 0.04% Poisoned Samples

Pengfei Xia,Yueqi Zeng,Ziqiang Li,Wei Zhang,Bin Li

This study focuses on reducing the number of poisoned samples needed when backdooring an image classifier. We present Efficient Trojan Injection (ETI), a pipeline that significantly improves the poisoning efficiency through trigger design, sample selection, and the exploitation of individual consistency. Using ETI, two backdoored datasets, CIFAR-10-B0-20 and CIFAR-100-B0-30, are constructed and released, in which 0.04% (20/50,000) and 0.06% (30/50,000) of the train images are poisoned. Across 240 models with different network architectures and training hyperparameters, the average attack success rates on these two sets are 92.1% and

90.4%, respectively. These results indicate that it is feasible to inject a Trojan into an image classifier with only a few tens of poisoned samples, which is about an order of magnitude less than before.

****

Priors, Hierarchy, and Information Asymmetry for Skill Transfer in Reinforcement Learning

Sasha Salter,Kristian Hartikainen,Walter Goodwin,Ingmar Posner

The ability to discover behaviours from past experience and transfer them to new tasks is a hallmark of intelligent agents acting sample-efficiently in the real world. Equipping embodied reinforcement learners with the same ability may be crucial for their successful deployment in robotics. While hierarchical and KL-regularized reinforcement learning individually hold promise here, arguably a hybrid approach could combine their respective benefits. Key to these fields is the use of information asymmetry across architectural modules to bias which skills are learnt. While asymmetry choice has a large influence on transferability, existing methods base their choice primarily on intuition in a domain-independent, potentially sub-optimal, manner. In this paper, we theoretically and empirically show the crucial expressivity-transferability trade-off of skills across sequential tasks, controlled by information asymmetry. Given this insight, we introduce Attentive Priors for Expressive and Transferable Skills (APES), a hierarchical KL-regularized method, heavily benefiting from both priors and hierarchy. Unlike existing approaches, APES automates the choice of asymmetry by learning it in a data-driven, domain-dependent, way based on our expressivity-transferability theorems. Experiments over complex transfer domains of varying levels of extrapolation and sparsity, such as robot block stacking, demonstrate the criticality of the correct asymmetric choice, with APES drastically outperforming previous methods.

****

Self-Supervised Set Representation Learning for Unsupervised Meta-Learning

Dong Bok Lee,Seanie Lee,Kenji Kawaguchi,Yunji Kim,Jihwan Bang,Jung-Woo Ha,Sung Ju Hwang

Unsupervised meta-learning (UML) essentially shares the spirit of self-supervised learning (SSL) in that their goal aims at learning models without any human supervision so that the models can be adapted to downstream tasks. Further, the learning objective of self-supervised learning, which pulls positive pairs closer and repels negative pairs, also resembles metric-based meta-learning. Metric-based meta-learning is one of the most successful meta-learning methods, which learns to minimize the distance between representations from the same class.

One notable aspect of metric-based meta-learning, however, is that it is widely interpreted as a set-level problem since the inference of discriminative class prototypes (or set representations) from few examples is crucial for the performance of downstream tasks. Motivated by this, we propose Set-SimCLR, a novel self-supervised set representation learning framework for targeting UML problem. Specifically, our Set-SimCLR learns a set encoder on top of instance representations to maximize the agreement between two sets of augmented samples, which are generated by applying stochastic augmentations to a given image. We theoretically analyze how our proposed set representation learning can potentially improve the generalization performance at the meta-test. We also empirically validate its effectiveness on various benchmark datasets, showing that Set-SimCLR largely outperforms both UML and instance-level self-supervised learning baselines.

****

Neural Episodic Control with State Abstraction

Zhuo Li,Derui Zhu,Yujing Hu,Xiaofei Xie,Lei Ma,YAN ZHENG,Yan Song,Yingfeng Chen,Jianjun Zhao

Existing Deep Reinforcement Learning (DRL) algorithms suffer from sample inefficiency. Generally, episodic control-based approaches are solutions that leverage highly rewarded past experiences to improve sample efficiency of DRL algorithms. However, previous episodic control-based approaches fail to utilize the latent information from the historical behaviors (\eg, state transitions, topological similarities, \etc) and lack scalability during DRL training. This work introduce s Neural Episodic Control with State Abstraction (NECSA), a simple but effective state abstraction-based episodic control containing a more comprehensive episodic memory, a novel state evaluation, and a multi-step state analysis. We evaluate our approach to the MuJoCo and Atari tasks in OpenAI gym domains. The experimental results indicate that NECSA achieves higher sample efficiency than the state-of-the-art episodic control-based approaches. Our data and code are available at the project website\footnote{\url{https://sites.google.com/view/drl-necsa}}.

****

Minibatch Stochastic Three Points Method for Unconstrained Smooth Minimization
Soumia Boucherouite,Grigory Malinovsky,Peter Richtárik,El houcine Bergou
In this paper, we propose a new zero order optimization method called minibatch stochastic three points (MiSTP) method to solve an unconstrained minimization problem in a setting where only an approximation of the objective function evaluation is possible. It is based on the recently proposed stochastic three points (STP) method (Bergou et al., 2020). At each iteration, MiSTP generates a random search direction in a similar manner to STP, but chooses the next iterate based solely on the approximation of the objective function rather than its exact evaluations. We also analyze our method’s complexity in the nonconvex and convex cases and evaluate its performance on multiple machine learning tasks.

****

Multigraph Topology Design for Cross-Silo Federated Learning
Tuong Khanh Long Do,Binh Xuan Nguyen,Toan Tran,Erman Tjiputra,Quang D. Tran,Hien Nguyen,Vuong Pham,Anh Nguyen
Cross-silo federated learning utilizes a few hundred reliable data silos with high-speed access links to jointly train a model. While this approach becomes a popular setting in federated learning, designing a robust topology to reduce the training time is still an open problem.

In this paper, we present a new multigraph topology for cross-silo federated learning. We first construct the multigraph using the overlay graph. We then parse this multigraph into different simple graphs with isolated nodes. The existence of isolated nodes allows us to perform model aggregation without waiting for other nodes, hence reducing the training time. We further propose a new distributed learning algorithm to use with our multigraph topology. The intensive experiments on public datasets show that our proposed method significantly reduces the training time compared with recent state-of-the-art topologies while ensuring convergence and maintaining the accuracy.

****

Universal Speech Enhancement with Score-based Diffusion
Joan Serrà,Santiago Pascual,Jordi Pons,Recep Oguz Araz,Davide Scaini
Removing background noise from speech audio has been the subject of considerable effort, especially in recent years due to the rise of virtual communication and amateur recordings. Yet background noise is not the only unpleasant disturbance that can prevent intelligibility: reverb, clipping, codec artifacts, problematic equalization, limited bandwidth, or inconsistent loudness are equally disturbing and ubiquitous. In this work, we propose to consider the task of speech enhancement as a holistic endeavor, and present a universal speech enhancement system that tackles 55 different distortions at the same time. Our approach consists of a generative model that employs score-based diffusion, together with a multi-resolution conditioning network that performs enhancement with mixture density networks. We show that this approach significantly outperforms the state of the art in a subjective test performed by expert listeners. We also show that it achieves competitive objective scores with just 4-8 diffusion steps, despite not considering any particular strategy for fast sampling. We hope that both our methodology and technical contributions encourage researchers and practitioners to adopt a universal approach to speech enhancement, possibly framing it as a generative task.

****

Partial Advantage Estimator for Proximal Policy Optimization
Xiulei Song,Yizhao Jin,Gregory Slabaugh,Simon Lucas
Estimation of value in policy gradient methods is a fundamental problem. General

ized Advantage Estimation (GAE) is an exponentially-weighted estimator of an advantage function similar to TD(λ). It substantially reduces the variance of policy gradient estimates at the expense of bias. In practical applications, a truncated GAE is used due to the incompleteness of the trajectory, which results in a large bias during estimation. To address this challenge, instead of using the all truncated GAE, we propose to take a part of the calculated GAE for updates, which significantly reduces the bias due to the incomplete trajectory. We perform experiments in MuJoCo and μRTS to investigate the effect of different partial coefficient and sampling lengths. We show that our partial GAE approach yields better empirical results in both environments.

****

Critical Sampling for Robust Evolution Behavior Learning of Unknown Dynamical Systems

Ce Zhang,Siqi Wu,Kailiang Wu,Zhihai He

We study the following new and important problem: given an unknown dynamical system, what is the minimum number of samples needed for effective learning of its governing laws and accurate prediction of its future evolution behavior, and how to select these critical samples? In this work, we propose to explore this problem based on a design approach. Specifically, starting from a small initial set of samples, we adaptively discover and collect critical samples to achieve increasingly accurate learning of the system evolution. One central challenge here is that we do not know the network modeling error of the ground-truth system state, which is however needed for critical sampling. To address this challenge, we introduce a multi-step reciprocal prediction network where a forward evolution network and a backward evolution network are designed to learn and predict the temporal evolution behavior in the forward and backward time directions, respectively. Very interestingly, we find that the desired network modeling error is highly correlated with the multi-step reciprocal prediction error. More importantly, this multi-step reciprocal prediction error can be directly computed from the current system state without knowing the ground-truth or data statistics. This allows us to perform dynamic selection of critical samples from regions with high network modeling errors and develop an adaptive sampling-learning method for dynamical systems. To achieve accurate and robust learning from this small set of critical samples, we introduce a joint spatial-temporal evolution network which incorporates spatial dynamics modeling into the temporal evolution prediction for robust learning of the system evolution operator with few samples. Our extensive experimental results demonstrate that our proposed method is able to dramatically reduce the numbers of samples needed for effective learning and accurate prediction of evolution behaviors of unknown dynamical systems by up to hundreds of times, especially for high-dimensional dynamical systems.

****

Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems

Phillip Lippe,Sara Magliacane,Sindy Löwe,Yuki M Asano,Taco Cohen,Efstratios Gavves

Causal representation learning is the task of identifying the underlying causal variables and their relations from high-dimensional observations, such as images. Recent work has shown that one can reconstruct the causal variables from temporal sequences of observations under the assumption that there are no instantaneous causal relations between them. In practical applications, however, our measurement or frame rate might be slower than many of the causal effects. This effectively creates ``instantaneous'' effects and invalidates previous identifiability results. To address this issue, we propose iCITRIS, a causal representation learning method that allows for instantaneous effects in intervened temporal sequences when intervention targets can be observed, e.g., as actions of an agent. iCITRIS identifies the potentially multidimensional causal variables from temporal observations, while simultaneously using a differentiable causal discovery method to learn their causal graph. In experiments on three datasets of interactive systems, iCITRIS accurately identifies the causal variables and their causal graph.

****

Visual Imitation Learning with Patch Rewards

Minghuan Liu,Tairan He,Weinan Zhang,Shuicheng YAN,Zhongwen Xu

Visual imitation learning enables reinforcement learning agents to learn to behave from expert visual demonstrations such as videos or image sequences, without explicit, well-defined rewards.

Previous reseaches either adopt supervised learning techniques or induce simple and coarse scalar rewards from pixels, neglecting the dense information contained in the image demonstrations.

In this work, we propose to measure the expertise of various local regions of image samples, or called patches, and recover multi-dimensional patch rewards accordingly.

Patch reward is a more precise rewarding characterization that serves as fine-grained expertise measurement and visual explainability tool.

Specifically, we present Adversarial Imitation Learning with Patch Rewards (PatchAIL), which employs a patch-based discriminator to measure the expertise of different local parts from given images and provide patch rewards.

The patch-based knowledge is also used to regularize the aggregated reward and stabilize the training.

We evaluate our method on the standard pixel-based benchmark DeepMind Control Suite.

The experiment results have demonstrated that PatchAIL outperforms baseline methods and provides valuable interpretations for visual demonstrations.

****

Planning Immediate Landmarks of Targets for Model-Free Skill Transfer across Agents

Minghuan Liu,Zhengbang Zhu,Menghui Zhu,Yuzheng Zhuang,Weinan Zhang,Jianye HAO

In reinforcement learning applications, agents usually need to deal with various input/output features when specified with different state and action spaces by their developers or physical restrictions, indicating re-training from scratch and considerable sample inefficiency, especially when agents follow similar solution steps to achieve tasks.

In this paper, we aim to transfer pre-trained skills to alleviate the above challenge. Specifically, we propose PILoT, i.e., Planning Immediate Landmarks of Targets. PILoT utilizes the universal decoupled policy optimization to learn a goal-conditioned state planner; then, we distill a goal-planner to plan immediate landmarks in a model-free style that can be shared among different agents. In our experiments, we show the power of PILoT on various transferring challenges, including few-shot transferring across action spaces and dynamics, from low-dimensional vector states to image inputs, from simple robot to complicated morphology; and we also illustrate PILoT provides a zero-shot transfer solution from a simple 2D navigation task to the harder Ant-Maze task.

****

AdaDQH Optimizer: Evolving from Stochastic to Adaptive by Auto Switch of Precondition Matrix

Yun Yue,Zhiling Ye,Jiadi Jiang,Yongchao Liu,Ke Zhang

Adaptive optimizers (e.g., Adam) have achieved tremendous success in deep learning. The key component of the optimizer is the precondition matrix, which provides more gradient information and adjusts the step size of each gradient direction. Intuitively, the closer the precondition matrix approximates the Hessian, the faster convergence and better generalization the optimizer can achieve in terms of iterations. However, this performance improvement is usually accompanied by a huge increase in the amount of computation. In this paper, we propose a new optimizer called AdaDQH to achieve better generalization with acceptable computational overhead. The intuitions are the trade-off of the precondition matrix between computation time and approximation of Hessian, and the auto switch of the precondition matrix from Stochastic Gradient Descent (SGD) to the adaptive optimizer. We evaluate AdaDQH on public datasets of Computer Vision (CV), Natural Language Processing (NLP) and Recommendation Systems (RecSys). The experimental results reveal that, compared to the State-Of-The-Art (SOTA) optimizers, AdaDQH can ach ieve significantly better or highly competitive performance. Furthermore, we analyze how AdaDQH is able to auto switch from stochastic to adaptive and the actual effects in different scenes. The code is available in the supplemental material.

****

CodeT: Code Generation with Generated Tests

Bei Chen,Fengji Zhang,Anh Nguyen,Daoguang Zan,Zeqi Lin,Jian-Guang Lou,Weizhu Chen

The task of generating code solutions for a given programming problem can benefit from the use of pre-trained language models such as Codex, which can produce multiple diverse samples. However, a major challenge for this task is to select the most appropriate solution from the multiple samples generated by the pre-trained language models. A natural way to evaluate the quality and correctness of a code solution is to run it against a set of test cases, but the manual creation of such test cases is often costly and time-consuming. In this paper, we propose a novel method, CodeT, that leverages the same pre-trained language models to automatically generate test cases for the code samples, thus reducing the human effort and increasing the coverage of the test scenarios. CodeT then executes the code samples using the generated test cases, and performs a dual execution agreement, which considers both the consistency of the outputs against the generated test cases and the agreement of the outputs with other code samples. We conduct comprehensive experiments on four benchmarks, HumanEval, MBPP, APPS and CodeContests, using five different pre-trained language models with varying sizes and capabilities. Our results show that CodeT can significantly improve the performance of code solution selection over previous methods, achieving remarkable and consistent gains across different models and benchmarks. For instance, CodeT improves the pass@1 metric on HumanEval to 65.8%, which represents an absolute improvement of 18.8% over the code-davinci-002 model, and an absolute improvement of more than 20% over the previous state-of-the-art results.

****

Learning Specialized Activation Functions for Physics-informed Neural Networks

Honghui Wang,Lu Lu,Shiji Song,Gao Huang

At the heart of network architectures lie the non-linear activation functions, the choice of which affects the model optimization and task performance. In computer vision and natural language processing, the Rectified Linear Unit is widely adopted across different tasks. However, there is no such default choice of activation functions in the context of physics-informed neural networks (PINNs). It is observed that PINNs exhibit high sensitivity to activation functions due to the various characteristics of each physics system, which makes the choice of the suitable activation function for PINNs a critical issue. Existing works usually choose activation functions in an inefficient trial-and-error manner. To address this problem, we propose to search automatically for the optimal activation function when solving different PDEs. This is achieved by learning an adaptive activation function as linear combinations of a set of candidate functions, whose coefficients can be directly optimized by gradient descent. In addition to its efficient optimization, the proposed method enables the discovery of novel activation function and the incorporation with prior knowledge about the PDE system. We can further enhance its search space with adaptive slope. The effectiveness of the proposed adaptive activation function is demonstrated on a series of benchmarks, including the Poisson's equation, Burgers' equation, Allen-Cahn equation, convection equation, Korteweg–de Vries equation and Cahn-Hilliard equation. The performance gain of the proposed method is further interpreted from the neural tangent kernel perspective. Code will be released.

****

Dateformer: Transformer Extends Look-back Horizon to Predict Longer-term Time Series

Julong Young,Junhui Chen,Feihu Huang,Jian Peng

Transformers have demonstrated impressive strength in long-term series forecasting. Existing prediction research mostly focused on mapping past short sub-series (lookback window) to future series (forecast window). The longer training datas et time series will be discarded, once training is completed. Models can merely rely on lookback window information for inference, which impedes models from analyzing time series from a global perspective. And these windows used by Transformers are quite narrow because they must model each time-step therein. Under this point-wise processing style, broadening windows will rapidly exhaust their model capacity. This, for fine-grained time series, leads to a bottleneck in information input and prediction output, which is mortal to long-term series forecasting. To overcome the barrier, we propose a brand-new methodology to use Transformer for time series prediction. Specifically, we split time series into patches by day and reform point-wise to patch-wise processing, which considerably enhances the information input and output of Transformers. To further help models leverage the whole training set's global information during inference, we distill the information, store it in time representations, and replace series with time representations as the main modeling entities. Our designed time-modeling Transformer---Dateformer yields state-of-the-art accuracy on 7 real-world datasets with a 33.6% relative improvement and extends the maximum forecast range to half-year.

****

CAMVR: Context-Adaptive Multi-View Representation Learning for Dense Retrieval

zhilin liao,XIANGTING HOU,Dongfang Lou,Ningyu Zhang,Huajun Chen

The recently proposed MVR (Multi-View Representation) model achieves remarkable performance in open-domain dense retrieval. In MVR, the document can match with multi-view queries by encoding the document into multiple representations. However, these representations tend to collapse into the same one when the percentage of documents answering multiple queries in training data is low. In this paper, we propose a CAMVR (Context-Adaptive Multi-View Representation) learning framework, which explicitly avoids the collapse problem by aligning each viewer token with different document snippets. In CAMVR, each viewer token is placed before each snippet to capture the local and global information with the consideration that answers of different view queries may scatter in one document. In addition, the view of the snippet containing the answer is used to explicitly supervise the learning process, from which the interpretability of view representation is provided. The extensive experiments show that CAMVR outperforms the existing models and achieves state-of-the-art results.

****

BIL: Bandit Inference Learning for Online Representational Similarity Test

Xiaoting Ji,Shuoxun Xu,Wenhai Cui,Linglong Kong,Xiaodong Yan

Similarity analysis is commonly used to determine the size of the discrepancy between two representations of a distribution pattern. In contrast to classical representational similarity analysis, which identifies disparate types of representations based on their shared similarity structures in distance matrices, this article proposes an online hypothesis testing procedure that will be able to determine whether a representation's difference from a constant is more significant than a predefined margin for streaming data. As a basic reinforcement learning model, two-armed bandits (TAB) are used to construct test statistics that update online. To achieve the most efficient testing results, an optimal strategy is developed for the TAB process. Asymptotic test statistics are discussed in theory, as are its corresponding explicit density functions, which are more accumulated than the normal distribution commonly applied in classical statistical analysis. Since the power of the proposed representative similarity test (RST) method is higher than that of the classical test, simulation studies support the validity of the proposed method.

****

Adam Accumulation to Reduce Memory Footprints of both Activations and Gradients for Large-scale DNN Training

Yijia Zhang,Yibo Han,Shijie Cao,Guohao Dai,Youshan Miao,Ting Cao,Fan Yang,Ningyi Xu

Running out of GPU memory has become a main bottleneck for large-scale DNN training. How to reduce the memory footprint during training has received intensive research attention. We find that previous gradient accumulation reduces activation memory but fails to be compatible with gradient memory reduction due to a cont radiction between preserving gradients and releasing gradients. To address this issue, we propose a novel optimizer accumulation method for Adam, named Adam Accumulation (AdamA), which enables reducing both activation and gradient memory. Specifically, AdamA directly integrates gradients into optimizer states and accumulates optimizer states over micro-batches, so that gradients can be released immediately after use. We mathematically and experimentally demonstrate AdamA yields the same convergence properties as Adam. Evaluated on transformer-based models, AdamA achieves up to 23% memory reduction compared to gradient accumulation with less than 2% degradation in training throughput. Notably, AdamA can work together with memory reduction methods for optimizer states to fit 1.26x~3.14x larger models over PyTorch and DeepSpeed baseline on GPUs with different memory capacities.

****

Learning to Generate Columns with Application to Vertex Coloring

Yuan Sun,Andreas T Ernst,Xiaodong Li,Jake Weiner

We present a new column generation approach based on Machine Learning (ML) for solving combinatorial optimization problems. The aim of our method is to generate high-quality columns that belong to an optimal integer solution, in contrast to the traditional approach that aims at solving linear programming relaxations. To achieve this aim, we design novel features to characterize a column, and develop an effective ML model to predict whether a column belongs to an optimal integer solution. We then use the ML model as a filter to select high-quality columns generated from a sampling method and use the selected columns to construct an integer solution. Our method is computationally fast compared to the traditional methods that generate columns by repeatedly solving a pricing problem. We demonstrate the efficacy of our method on the vertex coloring problem, by empirically showing that the columns selected by our ML model are significantly better, in terms of the integer solution that can be constructed from them, than those selected randomly or based only on their reduced cost. Further, we show that the columns generated by our method can be used as a warm start to boost the performance of a column generation-based heuristic.

****

TPC-NAS: Sub-Five-Minute Neural Architecture Search for Image Classification, Object-Detection, and Super-Resolution

Ming-shan Huang,Tzi-Dar Chiueh

Neural network models have become more sophisticated with the explosive development of AI and its applications. Automating the model search process is essential to explore a full range of neural architectures for satisfactory performance. However, most current NAS algorithms consume significant time and computing resources, and many cater only to image classification applications. This paper proposes the total path count (TPC) score, which requires only simple calculation based on the architecture information, as an efficient accuracy predictor. TPC score is not only simple to come by but also very effective. The Kendall rank correlation coefficient of the TPC scores and the accuracies of 20 architectures for the CIFAR100 problem is as high as 0.87. This paper also proposes TPC-NAS, a zero-shot NAS method leveraging the novel TPC score. TPC-NAS requires no training and inference, and can complete a NAS task for Imagenet and other vision applications in less than five CPU minutes. Then, we apply TPC-NAS to image classification, object detection, and super-resolution applications for further validation. In image classification, TPC-NAS finds an architecture that achieves 76.4% top-1 accuracy in ImageNet with 355M FLOPs, outperforming other NAS solutions. Starting with yolov4-p5, TPC-NAS comes up with a high-performance model with at least 2% mAP improvement over other NAS algorithms’ results in object detection. Finally, in the super-resolution application, TPC-NAS discovers a model with fewer than 300K parameters and generates images with 32.09dB PSNR in the Urban100 dataset. These three experiments convince us that the TPC-NAS method can swiftly deliver high-quality CNN architectures in diverse applications. The related source code is available at https://github.com/TPC-NAS/TPC.

****

The Role of ImageNet Classes in Fréchet Inception Distance

Tuomas Kynkäänniemi,Tero Karras,Miika Aittala,Timo Aila,Jaakko Lehtinen

Fréchet Inception Distance (FID) is the primary metric for ranking models in data-driven generative modeling. While remarkably successful, the metric is known to sometimes disagree with human judgement. We investigate a root cause of these discrepancies, and visualize what FID "looks at" in generated images. We show that the feature space that FID is (typically) computed in is so close to the ImageNet classifications that aligning the histograms of Top-N classifications between sets of generated and real images can reduce FID substantially — without actually improving the quality of results. Thus, we conclude that FID is prone to intentional or accidental distortions. As a practical example of an accidental distortion, we discuss a case where an ImageNet pre-trained FastGAN achieves a FID comparable to StyleGAN2, while being worse in terms of human evaluation.

****

Diffusion Models Already Have A Semantic Latent Space

Mingi Kwon,Jaeseok Jeong,Youngjung Uh

Diffusion models achieve outstanding generative performance in various domains. Despite their great success, they lack semantic latent space which is essential for controlling the generative process. To address the problem, we propose asymmetric reverse process (Asyrp) which discovers the semantic latent space in frozen pretrained diffusion models. Our semantic latent space, named h-space, has nice properties for accommodating semantic image manipulation: homogeneity, linearity, robustness, and consistency across timesteps. In addition, we measure editing strength and quality deficiency of a generative process at timesteps to provide a principled design of the process for versatility and quality improvements. Our method is applicable to various architectures (DDPM++, iDDPM, and ADM) and datasets (CelebA-HQ, AFHQ-dog, LSUN-church, LSUN-bedroom, and METFACES).

****

Improving group robustness under noisy labels using predictive uncertainty

Dongpin Oh,Dae Lee,Jeunghyun Byun,Bonggun Shin

The standard empirical risk minimization (ERM) can underperform on certain minority groups (i.e., waterbirds in lands or landbirds in water) due to the spurious correlation between the input and its label. Several studies have improved the worst-group accuracy by focusing on the high-loss samples. The hypothesis behind this is that such high-loss samples are spurious-cue-free (SCF) samples. However, these approaches can be problematic since the high-loss samples may also be samples with noisy labels in the real-world scenarios. To resolve this issue, we utilize the predictive uncertainty of a model to improve the worst-group accuracy under noisy labels. To motivate this, we theoretically show that the high-uncertainty samples are the SCF samples in the binary classification problem. This theoretical result implies that the predictive uncertainty is an adequate indicator to identify SCF samples in a noisy label setting. Motivated from this, we propose a novel Entropy based Debiasing (END) framework that prevents models from learning the spurious cues while being robust to the noisy labels. In the END framework, we first train the \textit{identification model} to obtain the SCF samples from a training set using its predictive uncertainty. Then, another model is trained on the dataset augmented with an oversampled SCF set. The experimental results show that our END framework outperforms other strong baselines on several real-world benchmarks that consider both the noisy labels and the spurious-cues.

****

Mutual Information Regularized Offline Reinforcement Learning

Xiao Ma,Bingyi Kang,Zhongwen Xu,Min Lin,Shuicheng YAN

Offline reinforcement learning (RL) aims at learning an effective policy from offline datasets without active interactions with the environment. The major challenge of offline RL is the distribution shift that appears when out-of-distribution actions are queried, which makes the policy improvement direction biased by extrapolation errors. Most existing methods address this problem by penalizing the policy for deviating from the behavior policy during policy improvement or making conservative updates for value functions during policy evaluation. In this work, we propose a novel MISA framework to approach offline RL from the perspecti

ve of Mutual Information between States and Actions in the dataset by directly constraining the policy improvement direction. Intuitively, mutual information measures the mutual dependence of actions and states, which reflects how a behavior agent reacts to certain environment states during data collection. To effectively utilize this information to facilitate policy learning, MISA constructs lower bounds of mutual information parameterized by the policy and Q-values. We show that optimizing this lower bound is equivalent to maximizing the likelihood of a one-step improved policy on the offline dataset. In this way, we constrain the policy improvement direction to lie in the data manifold. The resulting algorithm simultaneously augments the policy evaluation and improvement by adding a mutual information regularization. MISA is a general offline RL framework that unifies conservative Q-learning (CQL) and behavior regularization methods (e.g., TD3+BC) as special cases. Our experiments show that MISA performs significantly better than existing methods and achieves new state-of-the-art on various tasks of the D4RL benchmark.
****
EVC: Towards Real-Time Neural Image Compression with Mask Decay
Wang Guo-Hua,Jiahao Li,Bin Li,Yan Lu
Neural image compression has surpassed state-of-the-art traditional codecs (H.266/VVC) for rate-distortion (RD) performance, but suffers from large complexity and separate models for different rate-distortion trade-offs. In this paper, we propose an Efficient single-model Variable-bit-rate Codec (EVC), which is able to run at 30 FPS with 768x512 input images and still outperforms VVC for the RD performance. By further reducing both encoder and decoder complexities, our small model even achieves 30 FPS with 1920x1080 input images. To bridge the performance gap between our different capacities models, we meticulously design the mask decay, which transforms the large model's parameters into the small model automatically. And a novel sparsity regularization loss is proposed to mitigate shortcomings of L_p regularization. Our algorithm significantly narrows the performance gap by 50% and 30% for our medium and small models, respectively. At last, we advocate the scalable encoder for neural image compression. The encoding complexity is dynamic to meet different latency requirements. We propose decaying the large encoder multiple times to reduce the residual representation progressively. Both mask decay and residual representation learning greatly improve the RD performance of our scalable encoder. Our code is at https://github.com/microsoft/DCVC.
****
Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model
Yinhuai Wang,Jiwen Yu,Jian Zhang
Most existing Image Restoration (IR) models are task-specific, which can not be generalized to different degradation operators. In this work, we propose the Denoising Diffusion Null-Space Model (DDNM), a novel zero-shot framework for arbitrary linear IR problems, including but not limited to image super-resolution, colorization, inpainting, compressed sensing, and deblurring. DDNM only needs a pre-trained off-the-shelf diffusion model as the generative prior, without any extra training or network modifications. By refining only the null-space contents during the reverse diffusion process, we can yield diverse results satisfying both data consistency and realness. We further propose an enhanced and robust version, dubbed DDNM+, to support noisy restoration and improve restoration quality for hard tasks. Our experiments on several IR tasks reveal that DDNM outperforms other state-of-the-art zero-shot IR methods. We also demonstrate that DDNM+ can solve complex real-world applications, e.g., old photo restoration.
****
Predicting Cellular Responses with Variational Causal Inference and Refined Relational Information
Yulun Wu,Rob Barton,Zichen Wang,Vassilis N. Ioannidis,Carlo De Donno,Layne C Price,Luis F. Voloch,George Karypis
Predicting the responses of a cell under perturbations may bring important benefits to drug discovery and personalized therapeutics. In this work, we propose a novel graph variational Bayesian causal inference framework to predict a cell's

gene expressions under counterfactual perturbations (perturbations that this cell did not factually receive), leveraging information representing biological knowledge in the form of gene regulatory networks (GRNs) to aid individualized cellular response predictions. Aiming at a data-adaptive GRN, we also developed an adjacency matrix updating technique for graph convolutional networks and used it to refine GRNs during pre-training, which generated more insights on gene relations and enhanced model performance. Additionally, we propose a robust estimator within our framework for the asymptotically efficient estimation of marginal perturbation effect, which is yet to be carried out in previous works. With extensive experiments, we exhibited the advantage of our approach over state-of-the-art deep learning models for individual response prediction.

****

Exploit Unlabeled Data on the Server! Federated Learning via Uncertainty-aware Ensemble Distillation and Self-Supervision

Jae-Min Park,Won-jun Jang,Tae-Hyun Oh,Si-Hyeon Lee

Federated Learning (FL) is a distributed machine learning paradigm that involves the cooperation of multiple clients to train a server model. In practice, it is hard to assume that each client possesses large-scale data or many clients are always available to participate in FL for the same round, which may lead to data deficiency. This deficiency degrades the entire learning process. To resolve this challenge, we propose a Federated learning with entropy-weighted ensemble Distillation and Self-supervised learning (FedDS). FedDS reliably deals with situations where not only the amount of data per client but also the number of clients is scarce. This advantage is achieved by leveraging the prevalent unlabeled data in the server. We demonstrate the effectiveness of FedDS on classification tasks for CIFAR-10/100 and PathMNIST. In CIFAR-10, our method shows the improvement over FedAVG by 12.54% in data deficient regime, and by 17.16% and 23.56% in more challenging scenarios of noisy label or Byzantine client cases, respectively.

****

Sample Importance in SGD Training

Alessio Quercia,Hanno Scharr,Ira Assent

Deep learning requires increasingly bigger models and datasets to improve generalization on unseen data, where some training data samples may be more informative than others. We investigate this assumption in supervised image classification by biasing SGD (Stochastic Gradient Descent) to sample important samples more often during training of a classifier. In contrast to state-of-the-art, our approach does not require additional training iterations to estimate the sample importance, because it computes estimates once during training using the training prediction probabilities. In experiments, we see that our learning technique converges on par or faster in terms of training iterations and can achieve higher test accuracy compared to state-of-the-art, especially when datasets are not suitably balanced. Results suggest that sample importance has intrinsic balancing properties and that an importance weighted class distribution can converge faster than the usual balanced class distribution. Finally, in contrast to recent work, we find that sample importance is model dependent. Therefore, calculating sample importance during training, rather than in a pre-processing step, may be the only viable way to go.

****

ResAct: Reinforcing Long-term Engagement in Sequential Recommendation with Residual Actor

Wanqi Xue,Qingpeng Cai,Ruohan Zhan,Dong Zheng,Peng Jiang,Kun Gai,Bo An

Long-term engagement is preferred over immediate engagement in sequential recommendation as it directly affects product operational metrics such as daily active users (DAUs) and dwell time. Meanwhile, reinforcement learning (RL) is widely regarded as a promising framework for optimizing long-term engagement in sequential recommendation. However, due to expensive online interactions, it is very difficult for RL algorithms to perform state-action value estimation, exploration and feature extraction when optimizing long-term engagement. In this paper, we propose ResAct which seeks a policy that is close to, but better than, the online-serving policy. In this way, we can collect sufficient data near the learned pol icy so that state-action values can be properly estimated, and there is no need to perform online exploration. ResAct optimizes the policy by first reconstructing the online behaviors and then improving it via a Residual Actor. To extract long-term information, ResAct utilizes two information-theoretical regularizers to confirm the expressiveness and conciseness of features. We conduct experiments on a benchmark dataset and a large-scale industrial dataset which consists of tens of millions of recommendation requests. Experimental results show that our method significantly outperforms the state-of-the-art baselines in various long-term engagement optimization tasks.

****

Dataset Pruning: Reducing Training Data by Examining Generalization Influence

Shuo Yang,Zeke Xie,Hanyu Peng,Min Xu,Mingming Sun,Ping Li

The great success of deep learning heavily relies on increasingly larger training data, which comes at a price of huge computational and infrastructural costs. This poses crucial questions that, do all training data contribute to model's performance? How much does each individual training sample or a sub-training-set affect the model's generalization, and how to construct the smallest subset from the entire training data as a proxy training set without significantly sacrificing the model's performance? To answer these, we propose dataset pruning, an optimization-based sample selection method that can (1) examine the influence of removing a particular set of training samples on model's generalization ability with theoretical guarantee, and (2) construct the smallest subset of training data that yields strictly constrained generalization gap. The empirically observed generalization gap of dataset pruning is substantially consistent with our theoretical expectations. Furthermore, the proposed method prunes 40% training examples on the CIFAR-10 dataset, halves the convergence time with only 1.3% test accuracy decrease, which is superior to previous score-based sample selection methods.

****

Visual Timing For Sound Source Depth Estimation in the Wild

Wei Sun,Lili Qiu

Depth estimation enables a wide variety of 3D applications, such as robotics and autonomous driving. Despite significant work on various depth sensors, it is challenging to develop an all-in-one method to meet multiple basic criteria. In this paper, we propose a novel audio-visual learning scheme by integrating semantic features with physical spatial cues to boost monocular depth with only one microphone. Inspired by the flash-to-bang theory, we develop FBDepth, the first passive audio-visual depth estimation framework. It is based on the difference between the time-of-flight (ToF) of the light and the sound. We formulate sound source depth estimation as an audio-visual event localization task for collision events. To approach decimeter-level depth accuracy, we design a coarse-to-fine pipeline to push the temporary localization accuracy from event-level to millisecond-level by aligning audio-visual correspondence and manipulating optical flow. FBDepth feeds the estimated visual timestamp together with the audio clip and object visual features to regress the source depth. We use a mobile phone to collect 3.6K+ video clips with 24 different objects at up to 60m. FBDepth shows superior performance especially at a long range compared to monocular and stereo methods.

****

StrucTexTv2: Masked Visual-Textual Prediction for Document Image Pre-training

Yuechen Yu,Yulin Li,Chengquan Zhang,Xiaoqiang Zhang,Zengyuan Guo,Xiameng Qin,Kun Yao,Junyu Han,Errui Ding,Jingdong Wang

In this paper, we present StrucTexTv2, an effective document image pre-training framework, by performing masked visual-textual prediction. It consists of two self-supervised pre-training tasks: masked image modeling and masked language modeling, based on text region-level image masking. The proposed method randomly masks some image regions according to the bounding box coordinates of text words. The objectives of our pre-training tasks are reconstructing the pixels of masked image regions and the corresponding masked tokens simultaneously. Hence the pre-trained encoder can capture more textual semantics in comparison to the masked image modeling that usually predicts the masked image patches. Compared to the ma sked multi-modal modeling methods for document image understanding that rely on both the image and text modalities, StrucTexTv2 models image-only input and potentially deals with more application scenarios free from OCR pre-processing. Extensive experiments on mainstream benchmarks of document image understanding demonstrate the effectiveness of StrucTexTv2. It achieves competitive or even new state-of-the-art performance in various downstream tasks such as image classification, layout analysis, table structure recognition, document OCR, and information extraction under the end-to-end scenario.

****

Link Prediction without Graph Neural Networks

Zexi Huang,Mert Kosan,Arlei Lopes da Silva,Ambuj Singh

Link prediction, which consists of predicting edges based on graph features, is a fundamental task in many graph applications. As for several related problems, Graph Neural Networks (GNNs), which are based on an attribute-centric message-passing paradigm, have become the predominant framework for link prediction. GNNs have consistently outperformed traditional topology-based heuristics, but what contributes to their performance? Are there simpler approaches that achieve comparable or better results? To answer these questions, we first identify important limitations in how GNN-based link prediction methods handle the intrinsic class imbalance of the problem---due to the graph sparsity---in their training and evaluation. Moreover, we propose Gelato, a novel topology-centric framework that applies a topological heuristic to a graph enhanced by attribute information via graph learning. Our model is trained end-to-end with an N-pair loss on an unbiased training set to address class imbalance. Experiments show that Gelato is 145% more accurate, trains 11 times faster, infers 6,000 times faster, and has less than half of the trainable parameters compared to state-of-the-art GNNs for link prediction.

****

AdaStride: Using Adaptive Strides in Sequential Data for Effective Downsampling

Yoonhyung Lee,Kyomin Jung

The downsampling layer has been one of the most commonly used deep learning (DL) components in sequential data processing due to its several advantages. First, it improves the generalization performance of networks by acting as an information bottleneck, where it extracts task-relevant features and discards others. Second, it reduces data resolution allowing CNN layers to have larger receptive fields with smaller kernel sizes. Third, the reduced data resolution facilitates the use of Transformer networks in case of high-resolution data. Accordingly, there have been many studies on downsampling methods, but they have a limitation in that they apply the same downsampling ratio across a data instance. Using the same downsampling ratio uniformly for an entire data instance does not reflect the fact that the task-relevant information is not uniformly distributed in real data. In this paper, we introduce AdaStride, a downsampling method that can apply adaptively varying downsampling ratios across a sequential data instance given an overall downsampling ratio. Specifically, AdaStride learns to deploy adaptive strides in a sequential data instance. Therefore, it can preserve more information from task-relevant parts of a data instance by using smaller strides for those parts and larger strides for less relevant parts. To achieve this, we propose a novel training method called vector positioning that rearranges each time step of an input on a one-dimensional line segment without reordering, which is used to build an alignment matrix for the downsampling. In experiments conducted on three different tasks of audio classification, automatic speech recognition, and discrete representation learning, AdaStride outperforms other widely used standard downsampling methods showing its generality and effectiveness. In addition, we analyze how our AdaStride learns the effective adaptive strides to improve its performance in the tasks.

****

Subclass-balancing Contrastive Learning for Long-tailed Recognition

Chengkai Hou,Jieyu Zhang,Haonan Wang,Tianyi Zhou

Long-tailed recognition with imbalanced classes naturally emerges in practical machine learning applications. Existing methods such as data reweighing, resampli ng, and supervised contrastive learning enforce the class balance with a price of introducing imbalance between instances of head class and tail class, which may ignore the underlying rich semantic substructures of the former and exaggerate the biases in the latter. We overcome these drawbacks by a novel "subclass-balancing contrastive learning (SBCL)'' approach that clusters each head class into multiple subclasses of similar sizes as the tail classes and enforce representations to capture the two-layer class hierarchy between the original classes and their subclasses. Since the clustering is conducted in the representation space and updated during the course of training, the subclass labels preserve the semantic substructures of head classes. Meanwhile, it does not overemphasize tail class samples so each individual instance contribute to the representation learning equally. Hence, our method achieves both the instance- and subclass-balance, while the original class labels are also learned through contrastive learning among subclasses from different classes. We evaluate SBCL over a list of long-tailed benchmark datasets and it achieves the state-of-the-art performance. In addition, we present extensive analyses and ablation studies of SBCL to verify its advantages.

****

Effective Cross-instance Positive Relations for Generalized Category Discovery

Shaozhe Hao,Kai Han,Kwan-Yee K. Wong

We tackle the issue of generalized category discovery (GCD). GCD considers the open-world problem of automatically clustering a partially labelled dataset, in which the unlabelled data contain instances from novel categories and also the labelled classes. In this paper, we address the GCD problem without a known category number in the unlabelled data. We propose a framework, named CiP, to bootstrap the representation by exploiting Cross-instance Positive relations for contrastive learning in the partially labelled data which are neglected in existing methods. First, to obtain reliable cross-instance relations to facilitate the representation learning, we introduce a semi-supervised hierarchical clustering algorithm, named selective neighbor clustering (SNC), which can produce a clustering hierarchy directly from the connected components in the graph constructed by selective neighbors. We also extend SNC to be capable of label assignment for the unlabelled instances with the given class number. Moreover, we present a method to estimate the unknown class number using SNC with a joint reference score considering clustering indexes of both labelled and unlabelled data. Finally, we thoroughly evaluate our CiP framework on public generic image recognition datasets (CIFAR-10, CIFAR-100, and ImageNet-100) and challenging fine-grained datasets (CUB, Stanford Cars, and Herbarium19), all establishing the new state-of-the-art.

****

Plateau in Monotonic Linear Interpolation --- A "Biased" View of Loss Landscape for Deep Networks

Xiang Wang,Annie N. Wang,Mo Zhou,Rong Ge

Monotonic linear interpolation (MLI) --- on the line connecting a random initialization with the minimizer it converges to, the loss and accuracy are monotonic --- is a phenomenon that is commonly observed in the training of neural networks. Such a phenomenon may seem to suggest that optimization of neural networks is easy. In this paper, we show that the MLI property is not necessarily related to the hardness of optimization problems, and empirical observations on MLI for deep neural networks depend heavily on the biases. In particular, we show that interpolating both weights and biases linearly leads to very different influences on the final output, and when different classes have different last-layer biases on a deep network, there will be a long plateau in both the loss and accuracy interpolation (which existing theory of MLI cannot explain). We also show how the last-layer biases for different classes can be different even on a perfectly balanced dataset using a simple model. Empirically we demonstrate that similar intuitions hold on practical networks and realistic datasets.

****

Expected Gradients of Maxout Networks and Consequences to Parameter Initialization

Hanna Tseran,Guido Montufar

We study the gradients of a maxout network with respect to inputs and parameters and obtain bounds for the moments depending on the architecture and the parameter distribution. We observe that the distribution of the input-output Jacobian depends on the input, which complicates a stable parameter initialization. Based on the moments of the gradients, we formulate parameter initialization strategies that avoid vanishing and exploding gradients in wide networks. Experiments with deep fully-connected and convolutional networks show that this strategy improves SGD and Adam training of deep maxout networks. In addition, we obtain refined bounds on the expected number of linear regions, results on the expected curve length distortion, and results on the NTK.

****

The KFIoU Loss for Rotated Object Detection

Xue Yang,Yue Zhou,Gefan Zhang,Jirui Yang,Wentao Wang,Junchi Yan,XIAOPENG ZHANG,Qi Tian

Differing from the well-developed horizontal object detection area whereby the computing-friendly IoU based loss is readily adopted and well fits with the detection metrics, rotation detectors often involve a more complicated loss based on SkewIoU which is unfriendly to gradient-based training. In this paper, we propose an effective approximate SkewIoU loss based on Gaussian modeling and Gaussian product, which mainly consists of two items. The first term is a scale-insensitive center point loss, which is used to quickly narrow the distance between the center points of the two bounding boxes. In the distance-independent second term, the product of the Gaussian distributions is adopted to inherently mimic the mechanism of SkewIoU by its definition, and show its alignment with the SkewIoU loss at trend-level within a certain distance (i.e. within 9 pixels). This is in contrast to recent Gaussian modeling based rotation detectors e.g. GWD loss and KLD loss that involve a human-specified distribution distance metric which require additional hyperparameter tuning that vary across datasets and detectors. The resulting new loss called KFIoU loss is easier to implement and works better compared with exact SkewIoU loss, thanks to its full differentiability and ability to handle the non-overlapping cases. We further extend our technique to the 3-D case which also suffers from the same issues as 2-D. Extensive results on various datasets with different base detectors show the effectiveness of our approach.

****

Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting

Yunhao Zhang,Junchi Yan

Recently many deep models have been proposed for multivariate time series (MTS) forecasting. In particular, Transformer-based models have shown great potential because they can capture long-term dependency. However, existing Transformer-based models mainly focus on modeling the temporal dependency (cross-time dependency) yet often omit the dependency among different variables (cross-dimension dependency), which is critical for MTS forecasting. To fill the gap, we propose Crossformer, a Transformer-based model utilizing cross-dimension dependency for MTS forecasting. In Crossformer, the input MTS is embedded into a 2D vector array through the Dimension-Segment-Wise (DSW) embedding to preserve time and dimension information. Then the Two-Stage Attention (TSA) layer is proposed to efficiently capture the cross-time and cross-dimension dependency. Utilizing DSW embedding and TSA layer, Crossformer establishes a Hierarchical Encoder-Decoder (HED) to use the information at different scales for the final forecasting. Extensive experimental results on six real-world datasets show the effectiveness of Crossformer against previous state-of-the-arts.

****

Go-Explore with a guide: Speeding up search in sparse reward settings with goal-directed intrinsic rewards

Chong Min John Tan,Mehul Motani

Reinforcement Learning (RL) agents have traditionally been very sample-intensive to train, especially in environments with sparse rewards. Seeking inspiration from neuroscience experiments of rats learning the structure of a maze without ne eding extrinsic rewards, we seek to incorporate additional intrinsic rewards to guide behavior. We propose a potential-based goal-directed intrinsic reward (GDIR), which provides a reward signal regardless of whether the task is achieved, and ensures that learning can always take place. While GDIR may be similar to approaches such as reward shaping in incorporating goal-based rewards, we highlight that GDIR is innate to the agent and hence applicable across a wide range of environments without needing to rely on a properly shaped environment reward. We also note that GDIR is different from curiosity-based intrinsic motivation, which can diminish over time and lead to inefficient exploration. Go-Explore is a well-known state-of-the-art algorithm for sparse reward domains, and we demonstrate that by incorporating GDIR in the ``Go" and ``Explore" phases, we can improve Go-Explore's performance and enable it to learn faster across multiple environments, for both discrete (2D grid maze environments, Towers of Hanoi, Game of Nim) and continuous (Cart Pole and Mountain Car) state spaces. Furthermore, to consolidate learnt trajectories better, our method also incorporates a novel approach of hippocampal replay to update the values of GDIR and reset state visit and selection counts of states along the successful trajectory. As a benchmark, we also show that our proposed approaches learn significantly faster than traditional extrinsic-reward-based RL algorithms such as Proximal Policy Optimization, TD-learning, and Q-learning.

****

Is Stochastic Gradient Descent Near Optimal?

Yifan Zhu,Hong Jun Jeon,Benjamin Van Roy

The success of neural networks over the past decade has established them as effective models for many relevant data generating processes. Statistical theory on neural networks indicates graceful scaling of sample complexity. For example, Joen & Van Roy (An information-theoretic framework for supervised learning, 2022) demonstrate that, when data is generated by a ReLU teacher network with W parameters, an optimal learner needs only $\tilde{O}(W/\epsilon)$ samples to attain expected error ϵ. However, existing computational theory suggests that, even for single-hidden-layer teacher networks, to attain small error for all such teacher networks, the computation required to achieve this sample complexity is intractable. In this work, we fit single-hidden-layer neural networks to data generated by single-hidden-layer ReLU teacher networks with parameters drawn from a natural distribution. We demonstrate that stochastic gradient descent (SGD) with automated width selection attains small expected error with a number of samples and total number of queries both nearly linear in the input dimension and width. This suggests that SGD nearly achieves the information-theoretic sample complexity bounds of Joen & Van Roy (2022) in a computationally efficient manner. An important difference between our positive empirical results and the negative theoretical results is that the latter address worst-case error of deterministic algorithms, while our analysis centers on expected error of a stochastic algorithm.

****

BrainBERT: Self-supervised representation learning for intracranial recordings

Christopher Wang,Vighnesh Subramaniam,Adam Uri Yaari,Gabriel Kreiman,Boris Katz,Ignacio Cases,Andrei Barbu

We create a reusable Transformer, BrainBERT, for intracranial recordings bringing modern representation learning approaches to neuroscience. Much like in NLP and speech recognition, this Transformer enables classifying complex concepts, i.e., decoding neural data, with higher accuracy and with much less data by being pretrained in an unsupervised manner on a large corpus of unannotated neural recordings. Our approach generalizes to new subjects with electrodes in new positions and to unrelated tasks showing that the representations robustly disentangle the neural signal. Just like in NLP where one can study language by investigating what a language model learns, this approach opens the door to investigating the brain by what a model of the brain learns. As a first step along this path, we demonstrate a new analysis of the intrinsic dimensionality of the computations in different areas of the brain. To construct these representations, we combine a technique for producing super-resolution spectrograms of neural data with an ap proach designed for generating contextual representations of audio by masking. In the future, far more concepts will be decodable from neural recordings by using representation learning, potentially unlocking the brain like language models unlocked language.

****

Logic-aware Pre-training of Language Models

Siru Ouyang,Zhuosheng Zhang,hai zhao

Pre-trained language models (PrLMs) have been shown useful for enhancing a broad range of natural language understanding (NLU) tasks. However, the capacity for capturing logic relations in challenging NLU still remains a bottleneck even for state-of-the-art PrLM enhancement, which greatly stalls their reasoning abilities. To bridge the gap, we propose logic pre-training of language models to equip PrLMs with logical reasoning ability. To let logic pre-training perform on a clear, accurate, and generalized knowledge basis, we introduce \textit{fact} instead of the plain language unit in previous PrLMs. The \textit{fact} is extracted through syntactic parsing in avoidance of unnecessary complex knowledge injection. Meanwhile, it enables training logic-aware models to be conducted on a more general language text. To explicitly guide the PrLM to capture logic relations, three complementary self-supervised pre-training objectives are introduced: 1) logical structure completion to accurately capture fact-level logic from the original context, 2) logical path prediction on a logical graph to uncover global logic relationships among facts, 3) logical connectives masking to capture discourse-level for fact groups. We evaluate our model on a broad range of NLP tasks, including natural language inference, relation extraction, and machine reading comprehension with logical reasoning. Experimental results show that our model achieves significant performance in all the downstream tasks, especially in logical reasoning-related tasks.

****

Individual Fairness of Data Provider Regarding Privacy Risk and Gain

Toshiki Shibahara,Takayuki Miura,Masanobu Kii,Atsunori Ichikawa

Fairness and privacy risks are important concerns of machine learning (ML) when deploying ML to the real world. Recent studies have focused on group fairness and privacy protection, but no study focuses on individual fairness (IF) and privacy protection. In this paper, we propose a new definition of IF from the perspective of privacy protection and experimentally evaluate privacy-preserving ML based on the proposed IF. For the proposed definition, we assume that users provide their data to an ML service and consider the principle that all users should obtain gains corresponding to their privacy risks. As a user's gain, we calculate the accuracy improvement on the user's data when providing the data to the ML service. We conducted experiments on the image and tabular datasets using three neural networks (NNs) and two tree-based algorithms with differential privacy guarantee. The experimental results of NNs show that we cannot stably improve the proposed IF by changing the strength of privacy protection and applying defenses against membership inference attacks. The results of tree-based algorithms show that privacy risks were extremely small without depending on the strength of privacy protection but raise a new question about the motivation of users for providing their data.

****

Semi-connected Joint Entity Recognition and Relation Extraction of Contextual Entities in Family History Records

Daniel Segrera

Entity extraction is an important step in document understanding. Higher accuracy entity extraction on fine-grained entities can be achieved by combining the utility of Named Entity Recognition (NER) and Relation Extraction (RE) models. In this paper, a semi-connected joint model is proposed that implements NER and Relation extraction. This joint model utilizes relations between entities to infer context-dependent fine-grain named entities in text corpora. The RE module is prevented from conveying information to the NER module which reduces the error accumulation during training. That improves on the fine-grained NER F1-score of existing state-of-the-art from .4753 to .8563 on our data. This provides the potent ial for further applications in historical document processing. These applications will enable automated searching of historical documents, such as those used in economics research and family history.

****

On the Universality of Langevin Diffusion for Private Euclidean (Convex) Optimization

Arun Ganesh,Abhradeep Guha Thakurta,Jalaj Kumar Upadhyay

In this paper, we revisit the problem of differentially private empirical risk minimization (DP-ERM) and differentially private stochastic convex optimization (DP-SCO). We show that a well-studied continuous time algorithm from statistical physics, called Langevin diffusion (LD), simultaneously provides optimal privacy/utility trade-offs for both DP-ERM and DP-SCO, under ϵ-DP, and (ϵ,δ)-DP both for convex and strongly convex loss functions. We provide new time and dimension independent uniform stability properties of LD, with which we provide the corresponding optimal excess population risk guarantees for ϵ-DP. An important attribute of our DP-SCO guarantees for ϵ-DP is that they match the non-private optimal bounds as $\epsilon\to\infty$.

****

General Neural Gauge Fields

Fangneng Zhan,Lingjie Liu,Adam Kortylewski,Christian Theobalt

The recent advance of neural fields, such as neural radiance fields, has significantly pushed the boundary of scene representation learning. Aiming to boost the computation efficiency and rendering quality of 3D scenes, a popular line of research maps the 3D coordinate system to another measuring system, e.g., 2D manifolds and hash tables, for modeling neural fields. The conversion of coordinate systems can be typically dubbed as \emph{gauge transformation}, which is usually a pre-defined mapping function, e.g., orthogonal projection or spatial hash function. This begs a question: can we directly learn a desired gauge transformation along with the neural field in an end-to-end manner? In this work, we extend this problem to a general paradigm with a taxonomy of discrete and continuous cases, and develop an end-to-end learning framework to jointly optimize the gauge transformation and neural fields. To counter the problem that the learning of gauge transformations can collapse easily, we derive a general regularization mechanism from the principle of information conservation during the gauge transformation. To circumvent the high computation cost in gauge learning with regularization, we directly derive an information-invariant gauge transformation which allows to preserve scene information inherently and yield superior performance.

****

Learning Disentanglement in Autoencoders through Euler Encoding

Jaehoon Cha,Jeyan Thiyagalingam

Noting the importance of factorizing (or disentangling) the latent space, we propose a novel, non-probabilistic disentangling framework for autoencoders, based on the principles of symmetry transformations that are independent of one another. To the best of our knowledge, this is the first deterministic model that is aiming to achieve disentanglement based on autoencoders without pairs of images or labels, by explicitly introducing inductive biases into a model architecture through Euler encoding. The proposed model is then compared with a number of state-of-the-art models, relevant to disentanglement, including symmetry-based and generative models based on autoencoders. Our evaluation using six different disentanglement metrics, including the unsupervised disentanglement metric we propose here in this paper, shows that the proposed model can offer better disentanglement, especially when variances of the features are different, where other methods may struggle. We believe that this model opens several opportunities for linear disentangled representation learning based on deterministic autoencoders.

****

Nonlinear Reconstruction for Operator Learning of PDEs with Discontinuities

Samuel Lanthaler,Roberto Molinaro,Patrik Hadorn,Siddhartha Mishra

Discontinuous solutions arise in a large class of hyperbolic and advection-dominated PDEs. This paper investigates, both theoretically and empirically, the operator learning of PDEs with discontinuous solutions. We rigorously prove, in term s of lower approximation bounds, that methods which entail a linear reconstruction step (e.g. DeepONets or PCA-Nets) fail to efficiently approximate the solution operator of such PDEs. In contrast, we show that certain methods employing a non-linear reconstruction mechanism can overcome these fundamental lower bounds and approximate the underlying operator efficiently. The latter class includes Fourier Neural Operators and a novel extension of DeepONets termed shift-DeepONets. Our theoretical findings are confirmed by empirical results for advection equations, inviscid Burgers’ equation and the compressible Euler equations of gas dynamics.

****

Grafting Vision Transformers

Jongwoo Park,Kumara Kahatapitiya,Donghyun Kim,Shivchander Sudalairaj,Quanfu Fan,Michael S Ryoo

Vision Transformers (ViTs) have recently become the state-of-the-art across many computer vision tasks. In contrast to convolutional networks (CNNs), ViTs enable global information sharing even within shallow layers of a network, i.e., among high-resolution features. However, this perk was later overlooked with the success of pyramid architectures such as Swin Transformer, which show better performance-complexity trade-offs. In this paper, we present a simple and efficient add-on component (termed GrafT) that considers global dependencies and multi-scale information throughout the network, in both high- and low-resolution features alike. GrafT can be easily adopted in both homogeneous and pyramid Transformers while showing consistent gains. It has the flexibility of branching- out at arbitrary depths, widening a network with multiple scales. This grafting operation enables us to share most of the parameters and computations of the backbone, adding only minimal complexity, but with a higher yield. In fact, the process of progressively compounding multi-scale receptive fields in GrafT enables communications between local regions. We show the benefits of the proposed method on multiple benchmarks, including image classification (ImageNet-1K), semantic segmentation (ADE20K), object detection and instance segmentation (COCO2017). Our code and models will be made available.

****

Generate rather than Retrieve: Large Language Models are Strong Context Generators

Wenhao Yu,Dan Iter,Shuohang Wang,Yichong Xu,Mingxuan Ju,Soumya Sanyal,Chenguang Zhu,Michael Zeng,Meng Jiang

Knowledge-intensive tasks, such as open-domain question answering (QA), require access to a large amount of world or domain knowledge. A common approach for knowledge-intensive tasks is to employ a retrieve-then-read pipeline that first retrieves a handful of relevant contextual documents from an external corpus such as Wikipedia and then predicts an answer conditioned on the retrieved documents. In this paper, we present a novel perspective for solving knowledge-intensive tasks by replacing document retrievers with large language model generators. We call our method generate-then-read (GenRead), which first prompts a large language model to generate contextual documents based on a given question, and then reads the generated documents to produce the final answer. Furthermore, we propose a novel clustering-based prompting method that selects distinct prompts, in order to generate diverse documents that cover different perspectives, leading to better recall over acceptable answers. We conduct extensive experiments on three different knowledge-intensive tasks, including open-domain QA, fact checking, and dialogue system. Notably, GenRead achieves 71.6 and 54.4 exact match scores on TriviaQA and WebQ, significantly outperforming the state-of-the-art retrieve-then-read pipeline DPR-FiD by +4.0 and +3.9, without retrieving any documents from any external knowledge source. Lastly, we demonstrate the model performance can be further improved by combining retrieval and generation. Our code and generated documents can be found at https://github.com/wyu97/GenRead.

****

Online Continual Learning for Progressive Distribution Shift (OCL-PDS): A Practitioner's Perspective

Runtian Zhai,Stefan Schroedl,Aram Galstyan,Anoop Kumar,Greg Ver Steeg,Pradeep Na

tarajan
We introduce the novel OCL-PDS problem - Online Continual Learning for Progressive Distribution Shift. PDS refers to the subtle, gradual, and continuous distribution shift that widely exists in modern deep learning applications. It is widely observed in industry that PDS can cause significant performance drop. While previous work in continual learning and domain adaptation addresses this problem to some extent, our investigations from the practitioner's perspective reveal flawed assumptions that limit their applicability on daily challenges faced in real-world scenarios, and this work aims to close the gap between academic research and industry. For this new problem, we build 4 new benchmarks from the Wilds dataset, and implement 12 algorithms and baselines including both supervised and semi-supervised methods, which we test extensively on the new benchmarks. We hope that this work can provide practitioners with tools to better handle realistic PDS, and help scientists design better OCL algorithms.
****
Discovering Informative and Robust Positives for Video Domain Adaptation
Chang Liu,Kunpeng Li,Michael Stopa,Jun Amano,Yun Fu
Unsupervised domain adaptation for video recognition is challenging where the domain shift includes both spatial variations and temporal dynamics. Previous works have focused on exploring contrastive learning for cross-domain alignment. However, limited variations in intra-domain positives, false cross-domain positives, and false negatives hinder contrastive learning from fulfilling intra-domain discrimination and cross-domain closeness. This paper presents a non-contrastive learning framework without relying on negative samples for unsupervised video domain adaptation. To address the limited variations in intra-domain positives, we set unlabeled target videos as anchors and explored to mine "informative intra-domain positives" in the form of spatial/temporal augmentations and target nearest neighbors (NNs).

To tackle the false cross-domain positives led by noisy pseudo-labels, we reversely set source videos as anchors and sample the synthesized target videos as "robust cross-domain positives" from an estimated target distribution, which are naturally more robust to the pseudo-label noise. Our approach is demonstrated to be superior to state-of-the-art methods through extensive experiments on several cross-domain action recognition benchmarks.

****
Understanding Why Generalized Reweighting Does Not Improve Over ERM
Runtian Zhai,Chen Dan,J Zico Kolter,Pradeep Kumar Ravikumar
Empirical risk minimization (ERM) is known to be non-robust in practice to distributional shift where the training and the test distributions are different. A suite of approaches, such as importance weighting, and variants of distributionally robust optimization (DRO), have been proposed to solve this problem. But a line of recent work has empirically shown that these approaches do not significantly improve over ERM in real applications with distribution shift. The goal of this work is to obtain a comprehensive theoretical understanding of this intriguing phenomenon. We first posit the class of Generalized Reweighting (GRW) algorithms, as a broad category of approaches that iteratively update model parameters based on iterative reweighting of the training samples. We show that when overparameterized models are trained under GRW, the resulting models are close to that obtained by ERM. We also show that adding small regularization which does not greatly affect the empirical training accuracy does not help. Together, our results show that a broad category of what we term GRW approaches are not able to achieve distributionally robust generalization. Our work thus has the following sobering takeaway: to make progress towards distributionally robust generalization, we either have to develop non-GRW approaches, or perhaps devise novel classification/regression loss functions that are adapted to GRW approaches.
****
Betty: An Automatic Differentiation Library for Multilevel Optimization
Sang Keun Choe,Willie Neiswanger,Pengtao Xie,Eric Xing
Gradient-based multilevel optimization (MLO) has gained attention as a framework

for studying numerous problems, ranging from hyperparameter optimization and meta-learning to neural architecture search and reinforcement learning. However, gradients in MLO, which are obtained by composing best-response Jacobians via the chain rule, are notoriously difficult to implement and memory/compute intensive. We take an initial step towards closing this gap by introducing Betty, a software library for large-scale MLO. At its core, we devise a novel dataflow graph for MLO, which allows us to (1) develop efficient automatic differentiation for MLO that reduces the computational complexity from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2)$, (2) incorporate systems support such as mixed-precision and data-parallel training for scalability, and (3) facilitate implementation of MLO programs of arbitrary complexity while allowing a modular interface for diverse algorithmic and systems design choices. We empirically demonstrate that Betty can be used to implement an array of MLO programs, while also observing up to 11% increase in test accuracy, 14% decrease in GPU memory usage, and 20% decrease in training wall time over existing implementations on multiple benchmarks. We also showcase that Betty enables scaling MLO to models with hundreds of millions of parameters. We open-source the code at https://github.com/leopard-ai/betty.

****

PatchBlender: A Motion Prior for Video Transformers

Gabriele Prato,Yale Song,Janarthanan Rajendran,R Devon Hjelm,Neel Joshi,Sarath Chandar

Transformers have become one of the dominant architectures in the field of computer vision. However, there are yet several challenges when applying such architectures to video data. Most notably, these models struggle to model the temporal patterns of video data effectively. Directly targeting this issue, we introduce PatchBlender, a learnable blending function that operates over patch embeddings across the temporal dimension of the latent space. We show that our method is successful at enabling vision transformers to encode the temporal component of video data. On Something-Something v2 and MOVi-A, we show that our method improves the performance of a ViT-B. PatchBlender has the advantage of being compatible with almost any Transformer architecture and since it is learnable, the model can adaptively turn on or off the prior. It is also extremely lightweight compute-wise, 0.005% the GFLOPs of a ViT-B.

****

Linear Connectivity Reveals Generalization Strategies

Jeevesh Juneja,Rachit Bansal,Kyunghyun Cho,João Sedoc,Naomi Saphra

In the mode connectivity literature, it is widely accepted that there are common circumstances in which two neural networks, trained similarly on the same data, will maintain loss when interpolated in the weight space. In particular, transfer learning is presumed to ensure the necessary conditions for linear mode connectivity across training runs. In contrast to existing results from image classification, we find that among text classifiers (trained on MNLI, QQP, and CoLA), some pairs of finetuned models have large barriers of increasing loss on the linear paths between them. On each task, we find distinct clusters of models which are linearly connected on the test loss surface, but are disconnected from models outside the cluster---models that occupy separate basins on the surface. By measuring performance on specially-crafted diagnostic datasets, we find that these clusters correspond to different generalization strategies. For example, on MNLI, one cluster behaves like a bag of words model under domain shift, while another cluster uses syntactic heuristics. Our work demonstrates how the geometry of the loss surface can guide models towards different heuristic functions in standard finetuning settings.

****

CEREAL: Few-Sample Clustering Evaluation

Nihal V. Nayak,Ethan R. Elenberg,Clemens Rosenbaum

Evaluating clustering quality with reliable evaluation metrics like normalized mutual information (NMI) requires labeled data that can be expensive to annotate. We focus on the underexplored problem of estimating clustering quality with limited labels. We adapt existing approaches from the few-sample model evaluation literature to actively sub-sample, with a learned surrogate model, the most infor mative data points for annotation to estimate the evaluation metric. However, we find that their estimation can be biased and only relies on the labeled data. To that end, we introduce CEREAL, a comprehensive framework for few-sample clustering evaluation that extends active sampling approaches in three key ways. First, we propose novel NMI-based acquisition functions that account for the distinctive properties of clustering and uncertainties from a learned surrogate model. Next, we use ideas from semi-supervised learning and train the surrogate model with both the labeled and unlabeled data. Finally, we pseudo-label the unlabeled data with the surrogate model. We run experiments to estimate NMI in an active sampling pipeline on three datasets across vision and language. Our results show that CEREAL reduces the area under the absolute error curve by up to 57% compared to the best sampling baseline. We perform an extensive ablation study to show that our framework is agnostic to the choice of clustering algorithm and evaluation metric. We also extend CEREAL from clusterwise annotations to pairwise annotations. Overall, CEREAL can efficiently evaluate clustering with limited human annotations.

****

Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models

Meng Liu,Haoran Liu,Shuiwang Ji

Learning energy-based models (EBMs) is known to be difficult especially on discrete data where gradient-based learning strategies cannot be applied directly. Although ratio matching is a sound method to learn discrete EBMs, it suffers from expensive computation and excessive memory requirements, thereby resulting in difficulties in learning EBMs on high-dimensional data. Motivated by these limitations, in this study, we propose ratio matching with gradient-guided importance sampling (RMwGGIS). Particularly, we use the gradient of the energy function w.r.t. the discrete data space to approximately construct the provably optimal proposal distribution, which is subsequently used by importance sampling to efficiently estimate the original ratio matching objective. We perform experiments on density modeling over synthetic discrete data, graph generation, and training Ising models to evaluate our proposed method. The experimental results demonstrate that our method can significantly alleviate the limitations of ratio matching, perform more effectively in practice, and scale to high-dimensional problems. Our implementation is available at https://github.com/divelab/RMwGGIS.

****

Your Neighbors Are Communicating: Towards Powerful and Scalable Graph Neural Networks

Meng Liu,Haiyang Yu,Shuiwang Ji

Message passing graph neural networks (GNNs) are known to have their expressiveness upper-bounded by 1-dimensional Weisfeiler-Lehman (1-WL) algorithm. To achieve more powerful GNNs, existing attempts either require ad hoc features, or involve operations that incur high time and space complexities. In this work, we propose a general and provably powerful GNN framework that preserves the scalability of message passing scheme. In particular, we first propose to empower 1-WL for graph isomorphism test by considering edges among neighbors, giving rise to NC-1-WL. The expressiveness of NC-1-WL is shown to be strictly above 1-WL and below 3-WL theoretically. Further, we propose the NC-GNN framework as a differentiable neural version of NC-1-WL. Our simple implementation of NC-GNN is provably as powerful as NC-1-WL. Experiments demonstrate that our NC-GNN achieves remarkable performance on various benchmarks.

****

PATCorrect: Non-autoregressive Phoneme-augmented Transformer for ASR Error Correction

Ziji Zhang,Zhehui Wang,Rajesh Kamma,Sharanya Eswaran,Narayanan Sadagopan

Speech-to-text errors made by automatic speech recognition (ASR) system negatively impact downstream models relying on ASR transcriptions. Language error correction models as a post-processing text editing approach have been recently developed for refining the source sentences. However, efficient models for correcting errors in ASR transcriptions that meet the low latency requirements of industrial grade production systems have not been well studied. In this work, we propose a novel non-autoregressive (NAR) error correction approach to improve the transcription quality by reducing word error rate (WER) and achieve robust performance across different upstream ASR systems. Our approach augments the text encoding of the Transformer model with a phoneme encoder that embeds pronunciation information. The representations from phoneme encoder and text encoder are combined via a multi-modal fusion before feeding into the length tagging predictor for predicting target sequence lengths. The joint encoders also provide inputs to the attention mechanism in the NAR decoder. We experiment on 3 open-source ASR systems with varying speech-to-text transcription quality and their erroneous transcriptions on 2 public English corpus datasets. Results show that our PATCorrect (Phoneme Augmented Transformer for ASR error Correction) consistently outperforms state-of-the-art NAR error correction method on English corpus across different upstream ASR systems. For example, PATCorrect improves WER reduction by 20% compared to methods using text only modality and also achieves an inference latency comparable to other NAR models at tens of millisecond scale, especially on GPU hardware, while still being 4.2 - 6.7x times faster than autoregressive models on Common Voice and LibriSpeech datasets.

****

Composing Ensembles of Pre-trained Models via Iterative Consensus

Shuang Li,Yilun Du,Joshua B. Tenenbaum,Antonio Torralba,Igor Mordatch

Large pre-trained models exhibit distinct and complementary capabilities dependent on the data they are trained on. Language models such as GPT-3 are capable of textual reasoning but cannot understand visual information, while vision models such as DALL-E can generate photorealistic photos but fail to understand complex language descriptions. In this work, we propose a unified framework for composing ensembles of different pre-trained models -- combining the strengths of each individual model to solve various multimodal problems in a zero-shot manner. We use pre-trained models as "generators" or "scorers" and compose them via closed-loop iterative consensus optimization. The generator constructs proposals and the scorers iteratively provide feedback to refine the generated result. Such closed-loop communication enables models to correct errors caused by other models, significantly boosting performance on downstream tasks, e.g. improving accuracy on grade school math problems by 7.5%, without requiring any model finetuning. We demonstrate that consensus achieved by an ensemble of scorers outperforms the feedback of a single scorer, by leveraging the strengths of each expert model. Results show that the proposed method can be used as a general purpose framework for a wide range of zero-shot multimodal tasks, such as image generation, video question answering, mathematical reasoning, and robotic manipulation.

****

Automated Data Augmentations for Graph Classification

Youzhi Luo,Michael Curtis McThrow,Wing Yee Au,Tao Komikado,Kanji Uchino,Koji Maruhashi,Shuiwang Ji

Data augmentations are effective in improving the invariance of learning machines. We argue that the core challenge of data augmentations lies in designing data transformations that preserve labels. This is relatively straightforward for images, but much more challenging for graphs. In this work, we propose GraphAug, a novel automated data augmentation method aiming at computing label-invariant augmentations for graph classification. Instead of using uniform transformations as in existing studies, GraphAug uses an automated augmentation model to avoid compromising critical label-related information of the graph, thereby producing label-invariant augmentations at most times. To ensure label-invariance, we develop a training method based on reinforcement learning to maximize an estimated label-invariance probability. Experiments show that GraphAug outperforms previous graph augmentation methods on various graph classification tasks.

****

Learning Label Encodings for Deep Regression

Deval Shah,Tor M. Aamodt

Deep regression networks are widely used to tackle the problem of predicting a continuous value for a given input. Task-specialized approaches for training regr ession networks have shown significant improvement over generic approaches, such as direct regression. More recently, a generic approach based on regression by binary classification using binary-encoded labels has shown significant improvement over direct regression. The space of label encodings for regression is large. Lacking heretofore have been automated approaches to find a good label encoding for a given application. This paper introduces Regularized Label Encoding Learning (RLEL) for end-to-end training of an entire network and its label encoding. RLEL provides a generic approach for tackling regression. Underlying RLEL is our observation that the search space of label encodings can be constrained and efficiently explored by using a continuous search space of real-valued label encodings combined with a regularization function designed to encourage encodings with certain properties. These properties balance the probability of classification error in individual bits against error correction capability. Label encodings found by RLEL result in lower or comparable errors to manually designed label encodings. Applying RLEL results in 10.9% and 12.4% improvement in Mean Absolute Error (MAE) over direct regression and multiclass classification, respectively. Our evaluation demonstrates that RLEL can be combined with off-the-shelf feature extractors and is suitable across different architectures, datasets, and tasks. Code is available at https://github.com/ubc-aamodt-group/RLEL_regression.

****

Riemannian Metric Learning via Optimal Transport
Christopher Scarvelis,Justin Solomon

We introduce an optimal transport-based model for learning a metric tensor from cross-sectional samples of evolving probability measures on a common Riemannian manifold. We neurally parametrize the metric as a spatially-varying matrix field and efficiently optimize our model's objective using a simple alternating scheme. Using this learned metric, we can non-linearly interpolate between probability measures and compute geodesics on the manifold. We show that metrics learned using our method improve the quality of trajectory inference on scRNA and bird migration data at the cost of little additional cross-sectional data.

****

Computational-Unidentifiability in Representation for Fair Downstream Tasks
Taeuk Jang,Xiaoqian Wang

Deep representation learning methods are highlighted as they outperform classical algorithms in various downstream tasks, such as classification, clustering, generative models, etc. Due to their success and impact on the real world, fairness concern is rising with noticeable attention. However, the focus of the fairness problem was limited to a certain downstream task, mostly classification, and few were studied from the perspective of representation itself. We claim that the fairness problems to various downstream tasks originated from the input feature space, i.e., the learned representation space. While several studies explored fair representation for the classification task, the fair representation learning method for unsupervised learning is not actively discussed. To fill this gap, we define a new notion of fairness, computational-unidentifiability, which suggests the fairness of the representation as the distributional independence of the sensitive groups. We demonstrate motivating problems that achieving computationally-unidentifiable representation is critical for fair downstream tasks. Moreover, we propose a novel fairness metric, Fair Fréchet distance (FFD), to quantify the computational-unidentifiability and address the limitation of a well-known fairness metric for unsupervised learning, i.e., balance. The proposed metric is efficient in computation and preserves theoretical properties. We empirically validate the effectiveness of the computationally-unidentifiable representations in various downstream tasks.

****

Reliability of CKA as a Similarity Measure in Deep Learning
MohammadReza Davari,Stefan Horoi,Amine Natik,Guillaume Lajoie,Guy Wolf,Eugene Belilovsky

Comparing learned neural representations in neural networks is a challenging but important problem, which has been approached in different ways. The Centered Kernel Alignment (CKA) similarity metric, particularly its linear variant, has rec ently become a popular approach and has been widely used to compare representations of a network's different layers, of architecturally similar networks trained differently, or of models with different architectures trained on the same data. A wide variety of claims about similarity and dissimilarity of these various representations have been made using CKA results. In this work we present analysis that formally characterizes CKA sensitivity to a large class of simple transformations, which can naturally occur in the context of modern machine learning. This provides a concrete explanation to CKA sensitivity to outliers, which has been observed in past works, and to transformations that preserve the linear separability of the data, an important generalization attribute. We empirically investigate several weaknesses of the CKA similarity metric, demonstrating situations in which it gives unexpected or counterintuitive results. Finally we study approaches for modifying representations to maintain functional behaviour while changing the CKA value. Our results illustrate that, in many cases, the CKA value can be easily manipulated without substantial changes to the functional behaviour of the models, and call for caution when leveraging activation alignment metrics.

****

Gradient Properties of Hard Thresholding Operator
Saeed Damadi,Jinglai Shen

Sparse optimization receives increasing attention in many applications such as compressed sensing, variable selection in regression problems, and recently neural network compression in machine learning. For example, the problem of compressing a neural network is a bi-level, stochastic, and nonconvex problem that can be cast into a sparse optimization problem. Hence, developing efficient methods for sparse optimization plays a critical role in applications. The goal of this paper is to develop analytical techniques for general, large size sparse optimization problems using the hard thresholding operator. To this end, we study the iterative hard thresholding (IHT) algorithm, which has been extensively studied in the literature because it is scalable, fast, and easily implementable. In spite of extensive research on the IHT scheme, we develop several new techniques that not only recover many known results but also lead to new results. Specifically, we first establish a new and critical gradient descent property of the hard thresholding (HT) operator. Our gradient descent result can be related to the distance between points that are sparse. However, the distance between sparse points cannot provide any information about the gradient in the sparse setting. To the best of our knowledge, the other way around (the gradient to the distance) has not been shown so far in the literature. Also, our gradient descent property allows one to study the IHT when the stepsize is less than or equal to 1/L, where L>0 is the Lipschitz constant of the gradient of an objective function. Note that the existing techniques in the literature can only handle the case when the stepsize is strictly less than 1/L. By exploiting this we introduce and study HT-stable and HT-unstable stationary points and show no matter how close an initialization is to a HT-unstable stationary point (saddle point in sparse sense), the IHT sequence leaves it. Finally, we show that no matter what sparse initial point is selected, the IHT sequence converges if the function values at HT-stable stationary points are distinct, where the last condition is a new assumption that has not been found in the literature. We provide a video of 4000 independent runs where the IHT algorithm is initialized very close to a HT-unstable stationary point and show the sequences escape them.

****

Fair Attribute Completion on Graph with Missing Attributes
Dongliang Guo,Zhixuan Chu,Sheng Li

Tackling unfairness in graph learning models is a challenging task, as the unfairness issues on graphs involve both attributes and topological structures. Existing work on fair graph learning simply assumes that attributes of all nodes are available for model training and then makes fair predictions. In practice, however, the attributes of some nodes might not be accessible due to missing data or privacy concerns, which makes fair graph learning even more challenging. In this paper, we propose FairAC, a fair attribute completion method, to complement mis sing information and learn fair node embeddings for graphs with missing attributes. FairAC adopts an attention mechanism to deal with the attribute missing problem and meanwhile, it mitigates two types of unfairness, i.e., feature unfairness from attributes and topological unfairness due to attribute completion. FairAC can work on various types of homogeneous graphs and generate fair embeddings for them and thus can be applied to most downstream tasks to improve their fairness performance. To our best knowledge, FairAC is the first method that jointly addresses the graph attribution completion and graph unfairness problems. Experimental results on benchmark datasets show that our method achieves better fairness performance with less sacrifice in accuracy, compared with the state-of-the-art methods of fair graph learning. Code is available at: https://github.com/donglgcn/FairAC.

****

Comfort Zone: A Vicinal Distribution for Regression Problems

Roee Weiss Lifshitz,Omri Azencot

Domain-dependent data augmentation methods generate artificial samples using transformations suited for the underlying data domain, for example rotations on images and time warping on time series data. However, domain-independent approaches, e.g. mixup, are applicable to various data modalities, and as such they are general and versatile. While mixup-based techniques are used extensively in classification problems, their effect on regression tasks is somewhat less explored. To bridge this gap, we study the problem of domain-independent augmentation for regression, and we introduce comfort-zone: a new data-driven, domain-independent data augmentation method. Essentially, our approach samples new examples from the tangent planes of the train distribution. Augmenting data in this way aligns with the network tendency towards capturing the dominant features of its input signals. Evaluating comfort-zone on regression and time series forecasting benchmarks, we show that it improves the generalization of several neural architectures. We also find that mixup and noise injection are less effective in comparison to comfort-zone.

****

Implicit Neural Spatial Representations for Time-dependent PDEs

Honglin Chen,Rundi Wu,Eitan Grinspun,Changxi Zheng,Peter Yichen Chen

Numerically solving partial differential equations (PDEs) often entails spatial and temporal discretizations. Traditional methods (e.g., finite
difference, finite element, smoothed-particle hydrodynamics) frequently adopt explicit spatial discretizations, such as grids, meshes, and point clouds, where each degree-of-freedom corresponds to a location in space. While these explicit spatial correspondences are intuitive to model and understand, these representations are not necessarily optimal for accuracy, memory-usage, or adaptivity. In this work, we explore implicit neural representation as an alternative spatial discretization, where spatial information is implicitly stored in the neural network weights. With implicit neural spatial representation, PDE-constrained time-stepping translates into updating neural network weights, which naturally integrates with commonly adopted optimization time integrators. We validate our approach on a variety of classic PDEs with examples involving large elastic deformations, turbulent fluids, and multiscale phenomena. While slower to compute than traditional representations, our approach exhibits higher accuracy, lower memory consumption, and dynamically adaptive allocation of degrees of freedom without complex remeshing.

****

Deep Ranking Ensembles for Hyperparameter Optimization

Abdus Salam Khazi,Sebastian Pineda Arango,Josif Grabocka

Automatically optimizing the hyperparameters of Machine Learning algorithms is one of the primary open questions in AI. Existing work in Hyperparameter Optimization (HPO) trains surrogate models for approximating the response surface of hyperparameters as a regression task. In contrast, we hypothesize that the optimal strategy for training surrogates is to preserve the ranks of the performances of hyperparameter configurations as a Learning to Rank problem. As a result, we present a novel method that meta-learns neural network surrogates optimized for ra nking the configurations' performances while modeling their uncertainty via ensembling. In a large-scale experimental protocol comprising 12 baselines, 16 HPO search spaces and 86 datasets/tasks, we demonstrate that our method achieves new state-of-the-art results in HPO.

****

Multi-skill Mobile Manipulation for Object Rearrangement

Jiayuan Gu,Devendra Singh Chaplot,Hao Su,Jitendra Malik

We study a modular approach to tackle long-horizon mobile manipulation tasks for object rearrangement, which decomposes a full task into a sequence of subtasks. To tackle the entire task, prior work chains multiple stationary manipulation skills with a point-goal navigation skill, which are learned individually on subtasks. Although more effective than monolithic end-to-end RL policies, this framework suffers from compounding errors in skill chaining, e.g., navigating to a bad location where a stationary manipulation skill can not reach its target to manipulate. To this end, we propose that the manipulation skills should include mobility to have flexibility in interacting with the target object from multiple locations and at the same time the navigation skill could have multiple end points which lead to successful manipulation. We operationalize these ideas by implementing mobile manipulation skills rather than stationary ones and training a navigation skill trained with region goal instead of point goal. We evaluate our multi-skill mobile manipulation method M3 on 3 challenging long-horizon mobile manipulation tasks in the Home Assistant Benchmark (HAB), and show superior performance as compared to the baselines.

****

Robustness to corruption in pre-trained Bayesian neural networks

Xi Wang,Laurence Aitchison

We develop ShiftMatch, a new training-data-dependent likelihood for robustness to corruption in Bayesian neural networks (BNNs). ShiftMatch is inspired by the training-data-dependent “EmpCov” priors from Izmailov et al. (2021a), and efficiently matches test-time spatial correlations to those at training time. Critically, ShiftMatch is designed to leave the neural network’s training time likelihood unchanged, allowing it to use publicly available samples from pre-trained BNNs. Using pre-trained HMC samples, ShiftMatch gives strong performance improvements on CIFAR-10-C, outperforms EmpCov priors (though ShiftMatch uses extra information from a minibatch of corrupted test points), and is perhaps the first Bayesian method capable of convincingly outperforming plain deep ensembles.

****

Single-shot General Hyper-parameter Optimization for Federated Learning

Yi Zhou,Parikshit Ram,Theodoros Salonidis,Nathalie Baracaldo,Horst Samulowitz,Heiko Ludwig

We address the problem of hyper-parameter optimization (HPO) for federated learning (FL-HPO). We introduce Federated Loss SuRface Aggregation (FLoRA), a general FL-HPO solution framework that can address use cases of tabular data and any Machine Learning (ML) model including gradient boosting training algorithms, SVMs, neural networks, among others and thereby further expands the scope of FL-HPO. FLoRA enables single-shot FL-HPO: identifying a single set of good hyper-parameters that are subsequently used in a single FL training. Thus, it enables FL-HPO solutions with minimal additional communication overhead compared to FL training without HPO. Utilizing standard smoothness assumptions, we theoretically characterize the optimality gap of FLoRA for any convex and non-convex loss functions, which explicitly accounts for the heterogeneous nature of the parties' local data distributions, a dominant characteristic of FL systems. Our empirical evaluation of FLoRA for multiple FL algorithms on seven OpenML datasets demonstrates significant model accuracy improvements over the baselines, and robustness to increasing number of parties involved in FL-HPO training.

****

Weakly-supervised HOI Detection via Prior-guided Bi-level Representation Learning

Bo Wan,Yongfei Liu,Desen Zhou,Tinne Tuytelaars,Xuming He

Human object interaction (HOI) detection plays a crucial role in human-centric s

cene understanding and serves as a fundamental building block for many vision tasks. One generalizable and scalable strategy for HOI detection is to use weak supervision, learning from image-level annotations only. This is inherently challenging due to ambiguous human-object associations, large search space of detecting HOIs and highly noisy training signal. A promising strategy to address those challenges is to exploit knowledge from large-scale pretrained models (e.g., CLIP), but a direct knowledge distillation strategy does not perform well on the weakly-supervised setting. In contrast, we develop a CLIP-guided HOI representation capable of incorporating the prior knowledge at both image level and HOI instance level, and adopt a self-taught mechanism to prune incorrect human-object associations. Experimental results on HICO-DET and V-COCO
show that our method outperforms the previous works by a sizable margin, showing the efficacy of our HOI representation.
****

Meta-learning Adaptive Deep Kernel Gaussian Processes for Molecular Property Prediction

Wenlin Chen,Austin Tripp,José Miguel Hernández-Lobato

We propose Adaptive Deep Kernel Fitting with Implicit Function Theorem (ADKF-IFT), a novel framework for learning deep kernel Gaussian processes (GPs) by interpolating between meta-learning and conventional deep kernel learning. Our approach employs a bilevel optimization objective where we meta-learn generally useful feature representations across tasks, in the sense that task-specific GP models estimated on top of such features achieve the lowest possible predictive loss on average. We solve the resulting nested optimization problem using the implicit function theorem (IFT). We show that our ADKF-IFT framework contains previously proposed Deep Kernel Learning (DKL) and Deep Kernel Transfer (DKT) as special cases. Although ADKF-IFT is a completely general method, we argue that it is especially well-suited for drug discovery problems and demonstrate that it significantly outperforms previous state-of-the-art methods on a variety of real-world few-shot molecular property prediction tasks and out-of-domain molecular property prediction and optimization tasks.
****

ERL-Re^2: Efficient Evolutionary Reinforcement Learning with Shared State Representation and Individual Policy Representation

Jianye HAO,Pengyi Li,Hongyao Tang,YAN ZHENG,Xian Fu,Zhaopeng Meng

Deep Reinforcement Learning (Deep RL) and Evolutionary Algorithm (EA) are two major paradigms of policy optimization with distinct learning principles, i.e., gradient-based v.s. gradient-free. An appealing research direction is integrating Deep RL and EA to devise new methods by fusing their complementary advantages. However, existing works on combining Deep RL and EA have two common drawbacks:1) the RL agent and EA agents learn their policies individually, neglecting efficient sharing of useful common knowledge; 2) parameter-level policy optimization guarantees no semantic level of behavior evolution for the EA side. In this paper, we propose Evolutionary Reinforcement Learning with Two-scale State Representation and Policy Representation (ERL-Re^2), a novel solution to the aforementioned two drawbacks. The key idea of ERL-Re^2 is two-scale representation: all EA and RL policies share the same nonlinear state representation while maintaining individual linear policy representations. The state representation conveys expressive common features of the environment learned by all the agents collectively; the linear policy representation provides a favorable space for efficient policy optimization, where novel behavior-level crossover and mutation operations can be performed. Moreover, the linear policy representation allows convenient generalization of policy fitness with the help of Policy-extended Value Function Approximator (PeVFA), further improving the sample efficiency of fitness estimation. The experiments on a range of continuous control tasks show that ERL-Re^2 consistently outperforms advanced baselines and achieves the State Of The Art (SOTA). Our code is available on https://github.com/yeshenpy/ERL-Re2.
****

Least-to-Most Prompting Enables Complex Reasoning in Large Language Models

Denny Zhou,Nathanael Schärli,Le Hou,Jason Wei,Nathan Scales,Xuezhi Wang,Dale Sch

uurmans,Claire Cui,Olivier Bousquet,Quoc V Le,Ed H. Chi
Although chain-of-thought prompting has shown impressive results on many natural language reasoning tasks, it often performs poorly on tasks which need to solve problems harder than the demonstration examples. To tackle such easy-to-hard generalization issues, we propose a novel prompting strategy, least-to-most prompting. It is implemented through two stage prompting: reducing a complex problem into a list of subproblems, and then sequentially solving these subproblems, whereby solving a given subproblem is facilitated by the answers to previously solved subproblems. Experiments on symbolic manipulation, compositional generalization and math reasoning show that least-to-most prompting can generalize to the examples that are harder than those seen in the prompt, and outperform chain-of-thought prompting by a large margin. A notable result is that the GPT-3 code-davinci-002 model with least-to-most-prompting solves the SCAN benchmark regardless of splits (such as length split) with an accuracy of 99.7% using 14 examples versus an accuracy of 16.2% by chain-of-thought prompting, and neural-symbolic models in the literature specialized for solving SCAN are trained with the full training set of more than 15,000 examples.
****
Deep Ensembles for Graphs with Higher-order Dependencies
Steven Krieg,William Burgis,Patrick Soga,Nitesh Chawla
Graph neural networks (GNNs) continue to achieve state-of-the-art performance on many graph learning tasks, but rely on the assumption that a given graph is a sufficient approximation of the true neighborhood structure. In the presence of higher-order sequential dependencies, we show that the tendency of traditional graph representations to underfit each node's neighborhood causes existing GNNs to generalize poorly. To address this, we propose a novel Deep Graph Ensemble (DGE), which captures neighborhood variance by training an ensemble of GNNs on different neighborhood subspaces of the same node within a higher-order network structure. We show that DGE consistently outperforms existing GNNs on semisupervised and supervised tasks on six real-world data sets with known higher-order dependencies, even under a similar parameter budget. We demonstrate that learning diverse and accurate base classifiers is central to DGE's success, and discuss the implications of these findings for future work on GNNs.
****
Simplicial Embeddings in Self-Supervised Learning and Downstream Classification
Samuel Lavoie,Christos Tsirigotis,Max Schwarzer,Ankit Vani,Michael Noukhovitch,Kenji Kawaguchi,Aaron Courville
Simplicial Embeddings (SEM) are representations learned through self-supervised learning (SSL), wherein a representation is projected into L simplices of V dimensions each using a \texttt{softmax} operation. This procedure conditions the representation onto a constrained space during pretraining and imparts an inductive bias for group sparsity. For downstream classification, we formally prove that the SEM representation leads to better generalization than an unnormalized representation.
Furthermore, we empirically demonstrate that SSL methods trained with SEMs have improved generalization on natural image datasets such as CIFAR-100 and ImageNet. Finally, when used in a downstream classification task, we show that SEM features exhibit emergent semantic coherence where small groups of learned features are distinctly predictive of semantically-relevant classes.
****
Lossless Filter Pruning via Adaptive Clustering for Convolutional Neural Networks
Tao Niu,Yinglei Teng,Panpan Zou,Yiding Liu
The filter pruning method introduces structural sparsity by removing selected filters and is thus particularly effective for reducing complexity. However, previous works face two common limitations. 1) The pruned filters are prevented from contributing to the final outputs, resulting in performance degradation, especially when it comes to a large pruning rate. 2) To recover accuracy, the time-consuming fine-tuning step is required. The cost in time and the need for training data make it difficult to deploy in real-world scenarios. To address the aforemen

tioned limitations, we propose a novel filter pruning method called Cluster Pruning (CP). Our CP reconstructs the redundant filters from the perspective of similarity and removes them equivalently using the proposed channel addition operation in a lossless manner. Pruning in such a way allows CP to preserve as many learned features as possible while getting rid of the need for fine-tuning. Specifically, each filter is first distinguished by clustering and then reconstructed as the centroid to which it belongs. Filters are then updated to eliminate the effect caused by mistakenly selected. After convergence, CP can equivalently remove identical filters through the proposed channel addition operation. The strategies for adjusting the pruning rate and the adaptive coefficient for clustering make our CP even smoother and more efficient. Extensive experiments on CIFAR-10 and ImageNet datasets show that our method achieves the best trade-off between performance and complexity compared with other state-of-the-art algorithms.

****

Vision Transformer Adapter for Dense Predictions

Zhe Chen,Yuchen Duan,Wenhai Wang,Junjun He,Tong Lu,Jifeng Dai,Yu Qiao

This work investigates a simple yet powerful dense prediction task adapter for Vision Transformer (ViT). Unlike recently advanced variants that incorporate vision-specific inductive biases into their architectures, the plain ViT suffers inferior performance on dense predictions due to weak prior assumptions. To address this issue, we propose the ViT-Adapter, which allows plain ViT to achieve comparable performance to vision-specific transformers. Specifically, the backbone in our framework is a plain ViT that can learn powerful representations from large-scale multi-modal data. When transferring to downstream tasks, a pre-training-free adapter is used to introduce the image-related inductive biases into the model, making it suitable for these tasks. We verify ViT-Adapter on multiple dense prediction tasks, including object detection, instance segmentation, and semantic segmentation. Notably, without using extra detection data, our ViT-Adapter-L yields state-of-the-art 60.9 box AP and 53.0 mask AP on COCO test-dev. We hope that the ViT-Adapter could serve as an alternative for vision-specific transformers and facilitate future research. Code and models will be released at https://github.com/czczup/ViT-Adapter.

****

Towards Understanding Why Mask Reconstruction Pretraining Helps in Downstream Tasks

Jiachun Pan,Pan Zhou,Shuicheng YAN

For unsupervised pretraining, mask-reconstruction pretraining (MRP) approaches, e.g. MAE and data2vec, randomly mask input patches and then reconstruct the pixels or semantic features of these masked patches via an auto-encoder. Then for a downstream task, supervised fine-tuning the pretrained encoder remarkably surpasses the conventional "supervised learning" (SL) trained from scratch. However, it is still unclear 1) how MRP performs semantic (feature) learning in the pretraining phase and 2) why it helps in downstream tasks. To solve these problems, we first theoretically show that on an auto-encoder of a two/one-layered convolution encoder/decoder, MRP can capture all discriminative semantics of each potential semantic class in the pretraining dataset. Then considering the fact that the pretraining dataset is of huge size and high diversity and thus covers most semantics in downstream dataset, in fine-tuning phase, the pretrained encoder can capture as much semantics as it can in downstream datasets, and would not lost these semantics with theoretical guarantees. In contrast, SL only randomly captures some semantics due to lottery ticket hypothesis. So MRP provably achieves better performance than SL on the classification tasks.   Experimental results testify to our data assumptions and also our theoretical implications.

****

Similarity and Generalization: from Noise to Corruption

Veronica Guidetti,Nayara Fonseca

Contrastive learning aims to extract distinctive features from data by finding an embedding representation where similar samples are close to each other, and different ones are far apart. We study how NNs generalize the concept of similarity in the presence of noise, investigating two phenomena: Double Descent (DD) beh avior and online/offline correspondence. While DD examines how the network adjusts to the dataset during a long training time or by increasing the number of parameters, online/offline correspondence compares the network performances varying the quality (diversity) of the dataset. We focus on the simplest contrastive learning representative: Siamese Neural Networks (SNNs). We point out that SNNs can be affected by two distinct sources of noise: Pair Label Noise (PLN) and Single Label Noise (SLN). The effect of SLN is asymmetric, but it preserves similarity relations, while PLN is symmetric but breaks transitivity. We find that DD also appears in SNNs and is exacerbated by noise. We show that the dataset topology crucially affects generalization. While sparse datasets show the same performances under SLN and PLN for an equal amount of noise, SLN outperforms PLN in the overparametrized region in dense datasets. Indeed, in this regime, PLN similarity violation becomes macroscopical, corrupting the dataset to the point where complete overfitting cannot be achieved. We call this phenomenon Density-Induced Break of Similarity (DIBS). Probing the equivalence between online optimization and offline generalization in SNNs, we find that their correspondence breaks down in the presence of label noise for all the scenarios considered.

****

MEGAN: Multi Explanation Graph Attention Network

Jonas Teufel,Luca Torresi,Patrick Nicholas Reiser,Pascal Friederich

Explainable artificial intelligence (XAI) methods are expected to improve trust during human-AI interactions, provide tools for model analysis and extend human understanding of complex problems. Attention-based models are an important subclass of XAI methods, partly due to their full differentiability and the potential to improve explanations by means of explanation-supervised training. We propose the novel multi-explanation graph attention network (MEGAN). Our graph regression and classification model features multiple explanation channels, which can be chosen independently of the task specifications. We first validate our model on a synthetic graph regression dataset, where our model produces single-channel explanations with quality similar to GNNExplainer. Furthermore, we demonstrate the advantages of multi-channel explanations on one synthetic and two real-world datasets: The prediction of water solubility of molecular graphs and sentiment classification of movie reviews. We find that our model produces explanations consistent with human intuition, opening the way to learning from our model in less well-understood tasks.

****

Practical Approaches for Fair Learning with Multitype and Multivariate Sensitive Attributes

Tennison Liu,Alex Chan,Boris van Breugel,Mihaela van der Schaar

It is important to guarantee that machine learning algorithms deployed in the real world do not result in unfairness or unintended social consequences. Fair ML has largely focused on the protection of single attributes in the simpler setting where both attributes and target outcomes are binary. However, the practical application in many a real-world problem entails the simultaneous protection of multiple sensitive attributes, which are often not simply binary, but continuous or categorical. To address this more challenging task, we introduce FairCOCCO, a fairness measure built on cross-covariance operators on reproducing kernel Hilbert Spaces. This leads to two practical tools: first, the FairCOCCO Score, a normalized metric that can quantify fairness in settings with single or multiple sensitive attributes of arbitrary type; and second, a subsequent regularization term that can be incorporated into arbitrary learning objectives to obtain fair predictors. These contributions address crucial gaps in the algorithmic fairness literature, and we empirically demonstrate consistent improvements against state-of-the-art techniques in balancing predictive power and fairness on real-world datasets.

****

Self-Supervised Category-Level Articulated Object Pose Estimation with Part-Level SE(3) Equivariance

Xueyi Liu,Ji Zhang,Ruizhen Hu,Haibin Huang,He Wang,Li Yi

Category-level articulated object pose estimation aims to estimate a hierarchy o f articulation-aware object poses of an unseen articulated object from a known category. To reduce the heavy annotations needed for supervised learning methods, we present a novel self-supervised strategy that solves this problem without any human labels. Our key idea is to factorize canonical shapes and articulated object poses from input articulated shapes through part-level equivariant shape analysis. Specifically, we first introduce the concept of part-level SE(3) equivariance and devise a network to learn features of such property. Then, through a carefully designed fine-grained pose-shape disentanglement strategy, we expect that canonical spaces to support pose estimation could be induced automatically. Thus, we could further predict articulated object poses as per-part rigid transformations describing how parts transform from their canonical part spaces to the camera space. Extensive experiments demonstrate the effectiveness of our method on both complete and partial point clouds from synthetic and real articulated object datasets.

****

Universal Mini-Batch Consistency for Set Encoding Functions

Jeffrey Ryan Willette,Bruno Andreis,Juho Lee,Sung Ju Hwang

Previous works have established solid foundations for neural set functions, complete with architectures which preserve the necessary properties for operating on sets, such as invariance to permutations of the set elements. Subsequent work has highlighted the utility of Mini-Batch Consistency (MBC), the ability to sequentially process any permutation of a set partition scheme (e.g. streaming chunks of data) while maintaining consistency guarantees on the output, although there are limited options for MBC architectures. We propose a framework which can convert an arbitrary non-MBC model to one which satisfies MBC. In doing so, we allow all set functions to universally be considered in an MBC setting (UMBC). Additionally, we explore a Monte Carlo dropout strategy made possible by our framework which allows performing Monte Carlo dropout on streaming sets while never seeing the entire set at once. We validate UMBC with theoretical proofs, unit tests, and also provide qualitative/quantitative experiments on Gaussian data, clean and corrupted point cloud classification, and amortized clustering on ImageNet. Additionally, we investigate the probabilistic calibration of set-functions under test-time distributional shifts. Our results demonstrate the utility of universal mini-batch consistency, and we further discover that our dropout strategy improves uncertainty calibration.

****

Divide to Adapt: Mitigating Confirmation Bias for Domain Adaptation of Black-Box Predictors

Jianfei Yang,Xiangyu Peng,Kai Wang,Zheng Zhu,Jiashi Feng,Lihua Xie,Yang You

Domain Adaptation of Black-box Predictors (DABP) aims to learn a model on an unlabeled target domain supervised by a black-box predictor trained on a source domain. It does not require access to both the source-domain data and the predictor parameters, thus addressing the data privacy and portability issues of standard domain adaptation methods. Existing DABP approaches mostly rely on knowledge distillation (KD) from the black-box predictor, i.e., training the model with its noisy target-domain predictions, which however inevitably introduces the confirmation bias accumulated from the prediction noises and leads to degrading performance. To mitigate such bias, we propose a new strategy, \textit{divide-to-adapt}, that purifies cross-domain knowledge distillation by proper domain division. This is inspired by an observation we make for the first time in domain adaptation: the target domain usually contains easy-to-adapt and hard-to-adapt samples that have different levels of domain discrepancy w.r.t. the source domain, and deep models tend to fit easy-to-adapt samples first. Leveraging easy-to-adapt samples with less noise can help KD alleviate the negative effect of prediction noises from black-box predictors. In this sense, the target domain can be divided into an easy-to-adapt subdomain with less noise and a hard-to-adapt subdomain at the early stage of training. Then the adaptation is achieved by semi-supervised learning. We further reduce distribution discrepancy between subdomains and develop weak-strong augmentation strategy to filter the predictor errors progressively. As such, our method is a simple yet effective solution to reduce error accumul ation in cross-domain knowledge distillation for DABP. Moreover, we prove that the target error of DABP is bounded by the noise ratio of two subdomains, i.e., the confirmation bias, which provides the theoretical justifications for our method. Extensive experiments demonstrate our method achieves state of the art on all DABP benchmarks, outperforming the existing best approach by 7.0\% on VisDA-17, and is even comparable with the standard domain adaptation methods that use the source-domain data.

****

Thalamus: a brain-inspired algorithm for biologically-plausible continual learning and disentangled representations

Ali Hummos

Animals thrive in a constantly changing environment and leverage the temporal structure to learn well-factorized causal representations. In contrast, traditional neural networks suffer from forgetting in changing environments and many methods have been proposed to limit forgetting with different trade-offs. Inspired by the brain thalamocortical circuit, we introduce a simple algorithm that uses optimization at inference time to generate internal representations of the current task dynamically. The algorithm alternates between updating the model weights and a latent task embedding, allowing the agent to parse the stream of temporal experience into discrete events and organize learning about them. On a continual learning benchmark, it achieves competitive end average accuracy by mitigating forgetting, but importantly, the interaction between the weights dynamics and the latent dynamics organizes knowledge into flexible structures with a cognitive interface to control them. Tasks later in the sequence can be solved through knowledge transfer as they become reachable within the well-factorized latent space. The algorithm meets many of the desiderata of an ideal continually learning agent in open-ended environments, and its simplicity suggests fundamental computations in circuits with abundant feedback control loops such as the thalamocortical circuits in the brain

****

A Generalized EigenGame With Extensions to Deep Multiview Representation Learning

James William Harvey Chapman,Ana Lawry Aguila,Lennie Wells

Generalized Eigenvalue Problems (GEPs) encompass a range of interesting scientific computing problems. Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) are two such examples of GEPs which are often used to learn representations of multiview data. Development of efficient stochastic approaches to these problems would allow them to scale to large datasets. Furthermore, existing deep learning based extensions of CCA require large minibatch sizes in the stochastic setting to achieve good performance. Inspired by recent formulations of Principal Components Analysis and GEPs as games with differentiable utilities, we develop an alternative game theoretic approach to solving GEPs in which all constraints are softly enforced by Lagrange multipliers. We show that our approach shares much of the theoretical grounding of the previous game theoretic approaches but has fewer hyperparameters, is faster to converge, and permits extension to general function approximators like neural networks for certain GEPs including CCA. We demonstrate the effectiveness of our method for solving GEPs using canonical multiview datasets and demonstrate state-of-the-art performance for the Deep CCA problem for multiview representation learning.

****

Deep Variational Implicit Processes

Luis A. Ortega,Simon Rodriguez Santana,Daniel Hernández-Lobato

Implicit processes (IPs) are a generalization of Gaussian processes (GPs). IPs may lack a closed-form expression but are easy to sample from. Examples include, among others, Bayesian neural networks or neural samplers. IPs can be used as priors over functions, resulting in flexible models with well-calibrated prediction uncertainty estimates. Methods based on IPs usually carry out function-space approximate inference, which overcomes some of the difficulties of parameter-space approximate inference. Nevertheless, the approximations employed often limit the expressiveness of the final model, resulting, e.g., in a Gaussian predictive distribution, which can be restrictive. We propose here a multi-layer generalization of IPs called the Deep Variational Implicit process (DVIP). This generalization is similar to that of deep GPs over GPs, but it is more flexible due to the use of IPs as the prior distribution over the latent functions. We describe a scalable variational inference algorithm for training DVIP and show that it outperforms previous IP-based methods and also deep GPs. We support these claims via extensive regression and classification experiments. We also evaluate DVIP on large datasets with up to several million data instances to illustrate its good scalability and performance.

****

Denoising Masked Autoencoders Help Robust Classification

QuanLin Wu,Hang Ye,Yuntian Gu,Huishuai Zhang,Liwei Wang,Di He

In this paper, we propose a new self-supervised method, which is called denoising masked autoencoders (DMAE), for learning certified robust classifiers of images. In DMAE, we corrupt each image by adding Gaussian noises to each pixel value and randomly masking several patches. A Transformer-based encoder-decoder model is then trained to reconstruct the original image from the corrupted one. In this learning paradigm, the encoder will learn to capture relevant semantics for the downstream tasks, which is also robust to Gaussian additive noises. We show that the pre-trained encoder can naturally be used as the base classifier in Gaussian smoothed models, where we can analytically compute the certified radius for any data point. Although the proposed method is simple, it yields significant performance improvement in downstream classification tasks. We show that the DMAE ViT-Base model, which just uses 1/10 parameters of the model developed in recent work (Carlini et al., 2022), achieves competitive or better certified accuracy in various settings. The DMAE ViT-Large model significantly surpasses all previous results, establishing a new state-of-the-art on ImageNet dataset. We further demonstrate that the pre-trained model has good transferability to the CIFAR-10 dataset, suggesting its wide adaptability. Models and code are available at https://github.com/quanlin-wu/dmae.

****

Estimating individual treatment effects under unobserved confounding using binary instruments

Dennis Frauen,Stefan Feuerriegel

Estimating conditional average treatment effects (CATEs) from observational data is relevant in many fields such as personalized medicine. However, in practice, the treatment assignment is usually confounded by unobserved variables and thus introduces bias. A remedy to remove the bias is the use of instrumental variables (IVs). Such settings are widespread in medicine (e.g., trials where the treatment assignment is used as binary IV). In this paper, we propose a novel, multiply robust machine learning framework, called MRIV, for estimating CATEs using binary IVs and thus yield an unbiased CATE estimator. Different from previous work for binary IVs, our framework estimates the CATE directly via a pseudo-outcome regression. (1)~We provide a theoretical analysis where we show that our framework yields multiple robust convergence rates: our CATE estimator achieves fast convergence even if several nuisance estimators converge slowly. (2)~We further show that our framework asymptotically outperforms state-of-the-art plug-in IV methods for CATE estimation, in the sense that it achieves a faster rate of convergence if the CATE is smoother than the individual outcome surfaces. (3)~We build upon our theoretical results and propose a tailored deep neural network architecture called MRIV-Net for CATE estimation using binary IVs. Across various computational experiments, we demonstrate empirically that our MRIV-Net achieves state-of-the-art performance. To the best of our knowledge, our MRIV is the first multiply robust machine learning framework tailored to estimating CATEs in the binary IV setting.

****

Approximate Bayesian Inference with Stein Functional Variational Gradient Descent

Tobias Pielok,Bernd Bischl,David Rügamer

We propose a general-purpose variational algorithm that forms a natural analogue of Stein variational gradient descent (SVGD) in function space. While SVGD successively updates a set of particles to match a target density, the method introduced here of Stein functional variational gradient descent (SFVGD) updates a set of particle functions to match a target stochastic process (SP). The update step is found by minimizing the functional derivative of the Kullback-Leibler divergence between SPs. SFVGD can either be used to train Bayesian neural networks (BNNs) or for ensemble gradient boosting. We show the efficacy of training BNNs with SFVGD on various real-world datasets.

****

Soundness and Completeness: An Algorithmic Perspective on Evaluation of Feature Attribution

Yawei Li,Yang Zhang,Bernd Bischl,Mina Rezaei

Feature attribution is a fundamental approach to explaining neural networks by quantifying the importance of input features for a model's prediction. Although a variety of feature attribution methods have been proposed, there is little consensus on the assessment of attribution methods. In this study, we empirically show the limitations of \emph{order-based} and \emph{model-retraining} metrics. To overcome the limitations and enable evaluation with higher granularity, we propose a novel method to evaluate the \emph{completeness} and \emph{soundness} of feature attribution methods. Our proposed evaluation metrics are mathematically grounded on algorithm theory and require no knowledge of "ground truth" informative features. We validate our proposed metrics by conducting experiments on synthetic and real-world datasets. Lastly, we use the proposed metrics to benchmark a wide range of feature attribution methods. Our evaluation results provide an innovative perspective on comparing feature attribution methods. Code is in the supplementary material.

****

SCoMoE: Efficient Mixtures of Experts with Structured Communication

zhiyuan zeng,Deyi Xiong

Mixture-of-Experts (MoE) models are promising architectures for massively multilingual neural machine translation and large language models due to the advantage of sublinear scaling. However, the training of large MoE models is usually bottlenecked by the all-to-all communication (Lepikhin et al., 2020). To reduce the communication cost, we propose SCoMoE, an MoE architecture with structured all-to-all communication, inspired by the hierarchical architecture of the communication topology. SCoMoE encourages data to be communicated across devices through fast intra-accelerator/node communication channels, reducing communication throughput in the slow inter-node communication channel. We slice the data on the sequence dimension (SCoMoE-Seq) into three communication groups and project the data on the feature dimension (SCoMoE-Feat) into low-dimensional representations. To compensate the potential performance drop caused by the routing locality in SCoMoE, we further propose a token clustering approach to aggregating related tokens from different devices before the MoE layers. The sigmoid gating in the balanced router used in the token clustering is substituted with the softmax gating with differential sorting. Experiments on bilingual and massively multilingual machine translation demonstrate that SCoMoE achieves a speedup of 1.44x over GShard with comparable performance, and substantially outperforms Gshard (2.8 BLEU) on OPUS-100 with a speedup of 1.25x.

****

Locally Invariant Explanations: Towards Stable and Unidirectional Explanations through Local Invariant Learning

Amit Dhurandhar,Karthikeyan Natesan Ramamurthy,Kartik Ahuja,Vijay Arya

Locally interpretable model agnostic explanations (LIME) method is one of the most popular methods used to explain black-box models at a per example level. Although many variants have been proposed, few provide a simple way to produce high fidelity explanations that are also stable and intuitive. In this work, we provide a novel perspective by proposing a model agnostic local explanation method inspired by the invariant risk minimization (IRM) principle -- originally proposed for (global) out-of-distribution generalization -- to provide such high fidelity explanations that are also stable and unidirectional across nearby examples. O

ur method is based on a game theoretic formulation where we theoretically show that our approach has a strong tendency to eliminate features where the gradient of the black-box function abruptly changes sign in the locality of the example we want to explain, while in other cases it is more careful and will choose a more conservative (feature) attribution, a behavior which can be highly desirable for recourse. Empirically, we show on tabular, image and text data that the quality of our explanations with neighborhoods formed using random perturbations are much better than LIME and in some cases even comparable to other methods that use realistic neighbors sampled from the data manifold. This is desirable given that learning a manifold to either create realistic neighbors or to project explanations is typically expensive or may even be impossible. Moreover, our algorithm is simple and efficient to train, and can ascertain stable input features for local decisions of a black-box without access to side information such as a (partial) causal graph as has been seen in some recent works.
****

PLOT: Prompt Learning with Optimal Transport for Vision-Language Models
Guangyi Chen,Weiran Yao,Xiangchen Song,Xinyue Li,Yongming Rao,Kun Zhang
With the increasing attention to large vision-language models such as CLIP, there has been a significant amount of effort dedicated to building efficient prompts. Unlike conventional methods of only learning one single prompt, we propose to learn multiple comprehensive prompts to describe diverse characteristics of categories such as intrinsic attributes or extrinsic contexts. However, directly matching each prompt to the same visual feature is problematic, as it pushes the prompts to converge to one point. To solve this problem, we propose to apply optimal transport to match the vision and text modalities. Specifically, we first model images and the categories with visual and textual feature sets. Then, we apply a two-stage optimization strategy to learn the prompts. In the inner loop, we optimize the optimal transport distance to align visual features and prompts by the Sinkhorn algorithm, while in the outer loop, we learn the prompts by this distance from the supervised data. Extensive experiments are conducted on the few-shot recognition task and the improvement demonstrates the superiority of our method. The code is available at https://github.com/CHENGY12/PLOT.
****

An Additive Instance-Wise Approach to Multi-class Model Interpretation
Vy Vo,Van Nguyen,Trung Le,Quan Hung Tran,Reza Haf,Seyit Camtepe,Dinh Phung
Interpretable machine learning offers insights into what factors drive a certain prediction of a black-box system. A large number of interpreting methods focus on identifying explanatory input features, which generally fall into two main categories: attribution and selection. A popular attribution-based approach is to exploit local neighborhoods for learning instance-specific explainers in an additive manner. The process is thus inefficient and susceptible to poorly-conditioned samples. Meanwhile, many selection-based methods directly optimize local feature distributions in an instance-wise training framework, thereby being capable of leveraging global information from other inputs. However, they can only interpret single-class predictions and many suffer from inconsistency across different settings, due to a strict reliance on a pre-defined number of features selected. This work exploits the strengths of both methods and proposes a framework for learning local explanations simultaneously for multiple target classes. Our model explainer significantly outperforms additive and instance-wise counterparts on faithfulness with more compact and comprehensible explanations. We also demonstrate the capacity to select stable and important features through extensive experiments on various data sets and black-box model architectures.
****

Knowledge-Consistent Dialogue Generation with Language Models and Knowledge Graphs
Minki Kang,Jin Myung Kwak,Jinheon Baek,Sung Ju Hwang
Pre-trained language models have achieved impressive performances on dialogue generation tasks. However, when generating responses for a conversation that requires factual knowledge, they are far from perfect, due to the absence of mechanisms to retrieve, encode, and reflect the knowledge in the generated responses. So

me knowledge-grounded dialogue generation methods tackle this problem by leveraging the structured knowledge from Knowledge Graphs (KGs). However, existing methods do not guarantee that the model utilizes a relevant piece of knowledge from the KG before generating knowledge-consistent dialogues. To overcome this limitation, we propose SUbgraph Retrieval-augmented GEneration (SURGE), a framework for generating context-relevant and knowledge-consistent dialogues with a KG. Specifically, our method first retrieves the relevant subgraph from the KG, and then enforces consistency across facts by perturbing their word embeddings conditioned on the retrieved subgraph. Then, it learns a latent representation space using contrastive learning which ensures that the generated texts have high similarity to the retrieved subgraphs. We validate the performance of our SURGE framework on the OpendialKG and KOMODIS datasets and show that our method generates high-quality dialogues that faithfully reflect the knowledge from the KG.

****

Towards Semi-Supervised Learning with Non-Random Missing Labels

Yue Duan,Zhen Zhao,Lei Qi,Luping Zhou,Lei Wang,Yinghuan Shi

Semi-supervised learning (SSL) tackles the label missing problem by enabling the effective usage of unlabeled data. While existing SSL methods focus on the traditional setting, a practical and challenging scenario called label Missing Not At Random (MNAR) is usually ignored. In MNAR, the labeled and unlabeled data fall into different class distributions resulting in biased label imputation, which deteriorates the performance of SSL models. In this work, class transition tracking based Pseudo-Rectifying Guidance (PRG) is devised for MNAR. We explore the class-level guidance information obtained by the Markov random walk, which is modeled on a dynamically created graph built over the class tracking matrix. PRG unifies the history information of each class transition caused by the pseudo-rectifying procedure to activate the model's enthusiasm for neglected classes, so as the quality of pseudo-labels on both popular classes and rare classes in MNAR could be improved. We show the superior performance of PRG across a variety of the MNAR scenarios and the conventional SSL setting, outperforming the latest SSL solutions by a large margin. Checkpoints and evaluation code are available at the anonymous link https://anonymous.4open.science/r/PRG4SSL-MNAR-8DE2 while the source code will be available upon paper acceptance.

****

DASHA: Distributed Nonconvex Optimization with Communication Compression and Optimal Oracle Complexity

Alexander Tyurin,Peter Richtárik

We develop and analyze DASHA: a new family of methods for nonconvex distributed optimization problems. When the local functions at the nodes have a finite-sum or an expectation form, our new methods, DASHA-PAGE, DASHA-MVR and DASHA-SYNC-MVR, improve the theoretical oracle and communication complexity of the previous state-of-the-art method MARINA by Gorbunov et al. (2020). In particular, to achieve an $\varepsilon$-stationary point, and considering the random sparsifier Rand$K$ as an example, our methods compute the optimal number of gradients $\mathcal{O}\left(\frac{\sqrt{m}}{\varepsilon\sqrt{n}}\right)$ and $\mathcal{O}\left(\frac{\sigma}{\varepsilon^{3/2}n}\right)$ in finite-sum and expectation form cases, respectively, while maintaining the SOTA communication complexity $\mathcal{O}\left(\frac{d}{\varepsilon \sqrt{n}}\right)$. Furthermore, unlike MARINA, the new methods DASHA, DASHA-PAGE and DASHA-MVR send compressed vectors only, which makes them more practical for federated learning. We extend our results to the case when the functions satisfy the Polyak-Lojasiewicz condition. Finally, our theory is corroborated in practice: we see a significant improvement in experiments with nonconvex classification and training of deep learning models.

****

LDMIC: Learning-based Distributed Multi-view Image Coding

Xinjie Zhang,Jiawei Shao,Jun Zhang

Multi-view image compression plays a critical role in 3D-related applications. Existing methods adopt a predictive coding architecture, which requires joint encoding to compress the corresponding disparity as well as residual information. This demands collaboration among cameras and enforces the epipolar geometric cons traint between different views, which makes it challenging to deploy these methods in distributed camera systems with randomly overlapping fields of view. Meanwhile, distributed source coding theory indicates that efficient data compression of correlated sources can be achieved by independent encoding and joint decoding, which motivates us to design a learning-based distributed multi-view image coding (LDMIC) framework. With independent encoders, LDMIC introduces a simple yet effective joint context transfer module based on the cross-attention mechanism at the decoder to effectively capture the global inter-view correlations, which is insensitive to the geometric relationships between images. Experimental results show that LDMIC significantly outperforms both traditional and learning-based MIC methods while enjoying fast encoding speed. Code is released at https://github.com/Xinjie-Q/LDMIC.

****

Additive Poisson Process: Learning Intensity of Higher-Order Interaction in Poisson Processes

Simon Luo,Feng Zhou,lamiae azizi,Mahito Sugiyama

We present the Additive Poisson Process (APP), a novel framework that can model the higher-order interaction effects of the intensity functions in Poisson processes using projections into lower-dimensional space. Our model combines the techniques in information geometry to model higher-order interactions on a statistical manifold and in generalized additive models to use lower-dimensional projections to overcome the effects from the curse of dimensionality. Our approach solves a convex optimization problem by minimizing the KL divergence from a sample distribution in lower-dimensional projections to the distribution modeled by an intensity function in the Poisson process. Our empirical results show that our model is able to use samples observed in the lower dimensional space to estimate the higher-order intensity function with extremely sparse observations.

****

Sound Randomized Smoothing in Floating-Point Arithmetic

Vaclav Voracek,Matthias Hein

Randomized smoothing is sound when using infinite precision. However, we show that randomized smoothing is no longer sound for limited floating-point precision. We present a simple example where randomized smoothing certifies a radius of 1.26 around a point, even though there is an adversarial example in the distance 0.8 and show how this can be abused to give false certificates for CIFAR10. We discuss the implicit assumptions of randomized smoothing and show that they do not apply to generic image classification models whose smoothed versions are commonly certified. In order to overcome this problem, we propose a sound approach to randomized smoothing when using floating-point precision with essentially equal speed for quantized input. It yields sound certificates or image classifiers which for the ones tested so far are very similar to the unsound practice of randomized smoothing. Our only assumption is that we have access to a fair coin.

****

Assessing Model Out-of-distribution Generalization with Softmax Prediction Probability Baselines and A Correlation Method

Weijie Tu,Weijian Deng,Tom Gedeon,Liang Zheng

This paper studies the use of Softmax prediction to assess model generalization under distribution shift. Specifically, given an out-of distribution (OOD) test set and a pool of classifiers, we aim to develop a Softmax prediction-based measure which has a monotonic relationship with OOD generalization performance. We first show existing uncertainty measures (e.g., entropy and maximum Softmax prediction) are fairly useful of predicting generalization in some OOD scenarios. We then move ahead with proposing a new measure, Softmax Correlation (SoftmaxCorr). To obtain the SoftmaxCorr score for a classifier, we compute the class-class correlation matrix from all the Softmax vectors in a test set, and then its cosine similarity with an identity matrix. We show that the class-class correlation matrix reveals significant knowledge about the confusion matrix: its high similarity with the identity matrix means predictions have low confusion (uncertainty) and evenly cover all classes, and vice versa. Across three setups including Image

Net, CIFAR-10, and WILDS, we show that SoftmaxCorr is well predictive of model accuracy on both in-distribution and OOD datasets.
****

Collaborative Pure Exploration in Kernel Bandit

Yihan Du,Wei Chen,Yuko Kuroki,Longbo Huang

In this paper, we propose a novel Collaborative Pure Exploration in Kernel Bandit model (CoPE-KB), where multiple agents collaborate to complete different but related tasks with limited communication. Our model generalizes prior CoPE formulation with the single-task and classic MAB setting to allow multiple tasks and general reward structures. We propose a novel communication scheme with an efficient kernelized estimator, and design optimal algorithms CoKernelFC and CoKernelFB for CoPE-KB with fixed-confidence and fixed-budget objectives, respectively. Nearly matching upper and lower bounds in both sampling and communication complexity are established to demonstrate the optimality of our algorithms. Our theoretical results explicitly quantify how task similarities influence learning speedup, and only depend on the effective dimension of feature space. Our novel techniques including an efficient kernelized estimator and linear structured instance transformation, which overcome the communication difficulty in high-dimensional feature space and derive communication round lower bounds, can be of independent interests.
****

Provably Efficient Risk-Sensitive Reinforcement Learning: Iterated CVaR and Worst Path

Yihan Du,Siwei Wang,Longbo Huang

In this paper, we study a novel episodic risk-sensitive Reinforcement Learning (RL) problem, named Iterated CVaR RL, which aims to maximize the tail of the reward-to-go at each step, and focuses on tightly controlling the risk of getting into catastrophic situations at each stage. This formulation is applicable to real-world tasks that demand strong risk avoidance throughout the decision process, such as autonomous driving, clinical treatment planning and robotics. We investigate two performance metrics under Iterated CVaR RL, i.e., Regret Minimization and Best Policy Identification. For both metrics, we design efficient algorithms ICVaR-RM and ICVaR-BPI, respectively, and provide nearly matching upper and lower bounds with respect to the number of episodes K. We also investigate an interesting limiting case of Iterated CVaR RL, called Worst Path RL, where the objective becomes to maximize the minimum possible cumulative reward. For Worst Path RL, we propose an efficient algorithm with constant upper and lower bounds. Finally, the techniques we develop for bounding the change of CVaR due to the value function shift and decomposing the regret via a distorted visitation distribution are novel, and can find applications in other risk-sensitive online learning problems.
****

FedREP: A Byzantine-Robust, Communication-Efficient and Privacy-Preserving Framework for Federated Learning

Yi-Rui Yang,Kun Wang,Wu-Jun Li

Federated learning (FL) has recently become a hot research topic, in which Byzantine robustness, communication efficiency and privacy preservation are three important aspects. However, the tension among these three aspects makes it hard to simultaneously take all of them into account. In view of this challenge, we theoretically analyze the conditions that a communication compression method should satisfy to be compatible with existing Byzantine-robust methods and privacy-preserving methods. Motivated by the analysis results, we propose a novel communication compression method called consensus sparsification (ConSpar). To the best of our knowledge, ConSpar is the first communication compression method that is designed to be compatible with both Byzantine-robust methods and privacy-preserving methods. Based on ConSpar, we further propose a novel FL framework called FedREP, which is Byzantine-robust, communication-efficient and privacy-preserving. We theoretically prove the Byzantine robustness and the convergence of FedREP. Empirical results show that FedREP can significantly outperform communication-efficient privacy-preserving baselines. Furthermore, compared with Byzantine-robust communication-efficient baselines, FedREP can achieve comparable accuracy with an extra advantage of privacy preservation.

****

Few-Shot Transferable Robust Representation Learning via Bilevel Attacks

Hyeonjeong Ha,Minseon Kim,Sung Ju Hwang

Existing adversarial learning methods for enhancing the robustness of deep neural networks assume the availability of a large amount of data from which we can generate adversarial examples. However, in an adversarial meta-learning setting, the model needs to train with only a few adversarial examples to learn a robust model for unseen tasks, which is a very difficult goal to achieve. Further, learning transferable robust representations for unseen domains is a difficult problem even with a large amount of data. To tackle such a challenge, we propose a novel adversarial self-supervised meta-learning framework with bilevel attacks which aims to learn robust representations that can generalize across tasks and domains. Specifically, in the inner loop, we update the parameters of the given encoder by taking inner gradient steps using two different sets of augmented samples, and generate adversarial examples for each view by maximizing the instance classification loss. Then, in the outer loop, we meta-learn the encoder parameter to maximize the agreement between the two adversarial examples, which enables it to learn robust representations. We experimentally validate the effectiveness of our approach on unseen domain adaptation tasks, on which it achieves impressive performance. Specifically, our method significantly outperforms the state-of-the-art meta-adversarial learning methods on few-shot learning tasks, as well as self-supervised learning baselines in standard learning settings with large-scale datasets.

****

Targeted Adversarial Self-Supervised Learning

Minseon Kim,Hyeonjeong Ha,Sooel Son,Sung Ju Hwang

Recently, unsupervised adversarial training (AT) has been extensively studied to attain robustness with the models trained upon unlabeled data. To this end, previous studies have applied existing supervised adversarial training techniques to self-supervised learning (SSL) frameworks. However, all have resorted to untargeted adversarial learning as obtaining targeted adversarial examples is unclear in the SSL setting lacking of label information. In this paper, we propose a novel targeted adversarial training method for the SSL frameworks. Specifically, we propose a target selection algorithm for the adversarial SSL frameworks; it is designed to select the most confusing sample for each given instance based on similarity and entropy, and perturb the given instance toward the selected target sample. Our method significantly enhances the robustness of an SSL model without requiring large batches of images or additional models, unlike existing works aimed at achieving the same goal. Moreover, our method is readily applicable to general SSL frameworks that only uses positive pairs. We validate our method on benchmark datasets, on which it obtains superior robust accuracies, outperforming existing unsupervised adversarial training methods.

****

NIERT: Accurate Numerical Interpolation through Unifying Scattered Data Representations using Transformer Encoder

Shizhe Ding,Dongbo Bu

Numerical interpolation for scattered data, i.e., estimating values for target points based on those of some observed points, is widely used in computational science and engineering. The existing approaches either require explicitly pre-defined basis functions, which makes them inflexible and limits their performance in practical scenarios, or train neural networks as interpolators, which still have limited interpolation accuracy as they treat observed and target points separately and cannot effectively exploit the correlations among data points. Here, we present a learning-based approach to numerical interpolation for scattered data using encoder representation of Transformers (called NIERT). Unlike the recent learning-based approaches, NIERT treats observed and target points in a unified fashion through embedding them into the same representation space, thus gaining the advantage of effectively exploiting the correlations among them. The spec ially-designed partial self-attention mechanism used by NIERT makes it escape from the unexpected interference of target points on observed points. We further show that the partial self-attention is essentially a learnable interpolation module combining multiple neural basis functions, which provides interpretability of NIERT. Through pre-training on large-scale synthetic datasets, NIERT achieves considerable improvement in interpolation accuracy for practical tasks. On both synthetic and real-world datasets, NIERT outperforms the existing approaches, e.g., on the TFRD-ADlet dataset for temperature field reconstruction, NIERT achieves an MAE of 1.897×10^{-3}, substantially better than the state-of-the-art approach (MAE: 27.074×10^{-3}). The source code of NIERT is available at https://anonymous.4open.science/r/NIERT-2BCF.

****

Triplet Similarity Learning on Concordance Constraint

Jiansheng Fang,Jiajian Li,Jiang Liu

Triplet-based loss functions have been the paradigm of choice for robust deep metric learning (DML). However, conventional triplet-based losses require carefully tuning a decision boundary, i.e., violation margin. When performing online triplet mining on each mini-batch, choosing a good global and constant prior value for violation margin is challenging and irrational. To circumvent this issue, we propose a novel yet efficient concordance-induced triplet (CIT) loss as an objective function to train DML models. We formulate the similarity of triplet samples as a concordance constraint problem, then directly optimize concordance during DML model learning. Triplet concordance refers to the predicted ordering of intra-class and inter-class similarities being correct, which is invariant to any monotone transformation of the decision boundary of triplet samples. Hence, our CIT loss is free from the plague of adopting the violation margin as a prior constraint. In addition, due to the high training complexity of triplet-based losses, we introduce a partial likelihood term for CIT loss to impose additional penalties on hard triplet samples, thus enforcing fast convergence. We extensively experiment on a variety of DML tasks to demonstrate the elegance and simplicity of our CIT loss against its counterparts. In particular, on face recognition, person re-identification, as well as image retrieval datasets, our method can achieve comparable performances with state-of-the-arts without tuning any hyper-parameters laboriously.

****

Temporal Label Smoothing for Early Prediction of Adverse Events

Hugo Yèche,Alizée Pace,Gunnar Ratsch,Rita Kuznetsova

Models that can predict adverse events ahead of time with low false-alarm rates are critical to the acceptance of decision support systems in the medical community. This challenging machine learning task remains typically treated as a simple binary classification, with few bespoke methods proposed to leverage temporal dependency across samples. We propose Temporal Label Smoothing (TLS), a novel learning strategy that modulates smoothing strength as a function of proximity to the event of interest. This regularization technique reduces model confidence at the class boundary, where the signal is often noisy or uninformative, thus allowing training to focus on clinically informative data points away from this boundary region. From a theoretical perspective, we also show that our method can be framed as an extension of multi-horizon prediction, a learning heuristic proposed in other early prediction work. TLS empirically matches or outperforms all competitor methods across all evaluation measures on various early prediction benchmark tasks. In particular, our approach significantly improves performance on clinically-relevant metrics such as event recall under low false-alarm rates.

****

Test-Time Robust Personalization for Federated Learning

Liangze Jiang,Tao Lin

Federated Learning (FL) is a machine learning paradigm where many clients collaboratively learn a shared global model with decentralized training data. Personalization on FL models additionally adapts the global model to different clients, achieving promising results on consistent local training & test distributions. However, for real-world personalized FL applications, it is crucial to go one ste p further: robustifying FL models under the evolving local test set during deployment, where various types of distribution shifts can arise. In this work, we identify the pitfalls of existing works under test-time distribution shifts and propose Federated Test-time Head Ensemble plus tuning (FedTHE+), which personalizes FL models with robustness to various test-time distribution shifts. We illustrate the advancement of FedTHE+ (and its degraded computationally efficient variant FedTHE) over strong competitors, for training various neural architectures (CNN, ResNet, and Transformer) on CIFAR10 and ImageNet and evaluating on diverse test distributions. Along with this, we build a benchmark for assessing the performance and robustness of personalized FL methods during deployment. Code: \url{https://github.com/LINs-lab/FedTHE}.

****

What's Wrong with the Robustness of Object Detectors?

ZiYi Dong,Pengxu Wei,Liang Lin

Despite tremendous successes achieved, object detection models confront the vulnerability to adversarial attacks. Even with imperceptible adversarial perturbations in images, they probably yield erroneous detection predictions, posing a threat to various realistic applications, e.g., medical diagnosis and automatic driving. Although some existing methods can improve the adversarial robustness of detectors, they still suffer from the detection robustness bottleneck: the significant performance degradation on clean images and the limited robustness on adversarial images. In this paper, we conduct empirically a comprehensive investigation on what's wrong with the robustness of object detectors in four different seminal architectures, i.e., two-stage, one-stage, anchor-free, and Transformer-based detectors, inspiring more research interest on this task. We also devise a Detection Confusion Matrix (DCM) and Classification-Ablative Validation (ClsAVal) for further detection robustness analyses. We explore underlying factors that account for robustness bottleneck. It is empirically demonstrated that robust detectors have reliable localization robustness and poor classification robustness. The classification module easily mis-classifies the foreground objects into the background. Furthermore, Robust Derformable-DETR suffers from a poor classification and localization robustness. Our source codes, trained models, and detailed experiment results will be publicly available.

****

LAVA: Data Valuation without Pre-Specified Learning Algorithms

Hoang Anh Just,Feiyang Kang,Tianhao Wang,Yi Zeng,Myeongseob Ko,Ming Jin,Ruoxi Jia

Traditionally, data valuation is posed as a problem of equitably splitting the validation performance of a learning algorithm among the training data. As a result, the calculated data values depend on many design choices of the underlying learning algorithm. However, this dependence is undesirable for many use cases of data valuation, such as setting priorities over different data sources in a data acquisition process and informing pricing mechanisms in a data marketplace. In these scenarios, data needs to be valued before the actual analysis and the choice of the learning algorithm is still undetermined then. Another side-effect of the dependence is that to assess the value of individual points, one needs to re-run the learning algorithm with and without a point, which incurs a large computation burden.

This work leapfrogs over the current limits of data valuation methods by introducing a new framework that can value training data in a way that is oblivious to the downstream learning algorithm. Our main results are as follows. $\textbf{(1)}$ We develop a proxy for the validation performance associated with a training set based on a non-conventional $\textit{class-wise}$ $\textit{Wasserstein distance}$ between the training and the validation set. We show that the distance characterizes the upper bound of the validation performance for any given model under certain Lipschitz conditions. $\textbf{(2)}$ We develop a novel method to value individual data based on the sensitivity analysis of the $\textit{class-wise}$ Wasserstein distance. Importantly, these values can be directly obtained $\tex

tit{for free}$ from the output of off-the-shelf optimization solvers once the Wasserstein distance is computed. $\textbf{(3) }$We evaluate our new data valuation framework over various use cases related to detecting low-quality data
and show that, surprisingly, the learning-agnostic feature of our framework enables a $\textit{significant improvement}$ over the state-of-the-art performance while being $\textit{orders of magnitude faster.}$
****
FONDUE: an Algorithm to Find the Optimal Dimensionality of the Latent Representations of Variational Autoencoders
Lisa Bonheme,Marek Grzes
When training a variational autoencoder (VAE) on a given dataset, determining the optimal number of latent variables is mostly done by grid search: a costly process in terms of computational time and carbon footprint.
In this paper, we explore the intrinsic dimension estimation (IDE) of the data and latent representations learned by VAEs.
We show that the discrepancies between the IDE of the mean and sampled representations of a VAE after only a few steps of training reveal the presence of passive variables in the latent space, which, in well-behaved VAEs, indicates a superfluous number of dimensions.
Using this property, we propose FONDUE: an algorithm which quickly finds the number of latent dimensions after which the mean and sampled representations start to diverge (i.e., when passive variables are introduced), providing a principled method for selecting the number of latent dimensions for VAEs and autoencoders.
****
How do Variational Autoencoders Learn? Insights from Representational Similarity
Lisa Bonheme,Marek Grzes
The ability of Variational Autoencoders (VAEs) to learn disentangled representations has made them popular for practical applications. However, their behaviour is not yet fully understood. For example, the questions of when they can provide disentangled representations, or suffer from posterior collapse are still areas of active research. Despite this, there are no layerwise comparisons of the representations learned by VAEs, which would further our understanding of these models. In this paper, we thus look into the internal behaviour of VAEs using representational similarity techniques. Specifically, using the CKA and Procrustes similarities, we found that the encoders' representations are learned long before the decoders', and this behaviour is independent of hyperparameters, learning objectives, and datasets. Moreover, the encoders' representations in all but the mean and variance layers are similar across hyperparameters and learning objectives.
****
Meta-prediction Model for Distillation-Aware NAS on Unseen Datasets
Hayeon Lee,Sohyun An,Minseon Kim,Sung Ju Hwang
Distillation-aware Network Architecture Search (DaNAS) aims to search for an optimal student architecture that obtains the best performance and/or efficiency when distilling the knowledge from a given teacher model. Previous DaNAS methods have mostly tackled the search for the network architecture for fixed source/target tasks and the teacher, which are not generalized well on a new task, thus need to perform a costly search for any new combination of the domains and the teachers. For standard NAS tasks without KD, meta-learning-based computationally efficient NAS methods have been proposed, which learn the generalized search process over multiple tasks and transfer the knowledge obtained over those tasks to a new task. However, since they assume learning from scratch without KD from a teacher, they might not be ideal for DaNAS scenarios, which could significantly affect the final performances of the architectures obtained from the search. To eliminate the excessive computational cost of DaNAS methods and the sub-optimality of rapid NAS
methods, we propose a distillation-aware meta accuracy prediction model, DaSS (Distillation-aware Student Search), which can predict a given architecture’s final performances on a dataset when performing KD with a given teacher, without having actually to train it on the target task. The experimental results demonstrat

e that our proposed meta-prediction model successfully generalizes to multiple unseen datasets for DaNAS tasks, largely outperforming existing meta-NAS methods and rapid NAS baselines. Code is available at https://github.com/CownowAn/DaSS.

****

Manifold Characteristics That Predict Downstream Task Performance

Ruan Henry Van der Merwe,Gregory Newman,Etienne Barnard

Pretraining methods are typically compared by evaluating the accuracy of linear classifiers, transfer learning performance, or visually inspecting the representation manifold's (RM) lower-dimensional projections. We show that the differences between methods can be understood more clearly by investigating the RM directly, which allows for a more detailed comparison. To this end, we propose a framework and new metric to measure and compare different RMs. We also investigate and report on the RM characteristics for various pretraining methods. These characteristics are measured by applying sequentially larger local alterations to the input data, using white noise injections and Projected Gradient Descent (PGD) adversarial attacks, and then tracking each datapoint. We calculate the total distance moved for each datapoint and the relative change in distance between successive alterations. We show that self-supervised methods learn an RM where alterations lead to large but constant size changes, indicating a smoother RM than fully supervised methods. We then combine these measurements into one metric, the Representation Manifold Quality Metric (RMQM), where larger values indicate larger and less variable step sizes, and show that RMQM correlates positively with performance on downstream tasks.

****

Context Autoencoder for Self-Supervised Representation Learning

Xiaokang Chen,Mingyu Ding,Xiaodi Wang,Ying Xin,Shentong Mo,Yunhao Wang,Shumin Han,Ping Luo,Gang Zeng,Jingdong Wang

We present a novel masked image modeling (MIM) approach, context autoencoder (CAE), for self-supervised representation pretraining. The goal is to pretrain an encoder by solving the pretext task: estimate the masked patches from the visible patches in an image. Our approach first feeds the visible patches into the encoder, extracting the representations. Then, we make predictions from visible patches to masked patches in the encoded representation space. We introduce an alignment constraint, encouraging that the representations for masked patches, predicted from the encoded representations of visible patches, are aligned with the masked patch presentations computed from the encoder. In other words, the predicted representations are expected to lie in the encoded representation space, which empirically shows the benefit to representation learning. Last, the predicted masked patch representations are mapped to the targets of the pretext task through a decoder.

One additional characteristic is that our approach encourages the separation of the representation learning part (encoder), and the pretext task completion part that will be replaced by the downstream task part. In contrast, previous MIM methods (e.g., BEiT and MAE) couple the two parts, potentially limiting the representation learning quality. We demonstrate the effectiveness of our CAE through superior transfer performance in downstream tasks: semantic segmentation, and object detection and instance segmentation.

****

Learning to Linearize Deep Neural Networks for Secure and Efficient Private Inference

Souvik Kundu,Shunlin Lu,Yuke Zhang,Jacqueline Tiffany Liu,Peter Anthony Beerel

The large number of ReLU non-linearity operations in existing deep neural networks makes them ill-suited for latency-efficient private inference (PI). Existing techniques to reduce ReLU operations often involve manual effort and sacrifice significant accuracy. In this paper, we first present a novel measure of non-linearity layers’ ReLU sensitivity, enabling mitigation of the time-consuming manual efforts in identifying the same. Based on this sensitivity, we then present SENet, a three-stage training method that for a given ReLU budget, automatically assigns per-layer ReLU counts, decides the ReLU locations for each layer’s activation map, and trains a model with significantly fewer ReLUs to potentially yield latency and communication efficient PI. Experimental evaluations with multiple models on various datasets show SENet’s superior performance both in terms of reduced ReLUs and improved classification accuracy compared to existing alternatives. In particular, SENet can yield models that require up to ~2× fewer ReLUs while yielding similar accuracy. For a similar ReLU budget SENet can yield models with ~2.32% improved classification accuracy, evaluated on CIFAR-100.

****

Wasserstein Fair Autoencoders

Sungdong Lee,Hyunjong Lee,Joong-Ho Won

Autoencoders, or nonlinear factor models parameterized by neural networks, have become an indispensable tool for generative modeling and representation learning in high dimensions. Imposing structural constraints such as conditional independence on the latent variables (representation, or factors) in order to capture invariance or fairness with autoencoders has been attempted through adding ad hoc penalties to the loss function mostly in the variational autoencoder (VAE) context, often based on heuristic arguments. In this paper, we demonstrate that Wasserstein autoencoders (WAEs) are highly flexible in embracing structural constraints. Well-known extensions of VAEs for this purpose are gracefully handled within the framework of the seminal result by Tolstikhin et al. (2018). In particular, given a conditional independence structure of the generative model (decoder), corresponding encoder structure and penalties are induced from the functional constraints that define the WAE. This property of WAEs opens up a principled way of penalizing autoencoders to impose structural constraints. Utilizing this generative model structure, we present results on fair representation and conditional generation tasks, and compare them with other preceding methods.

****

Denoising Diffusion Error Correction Codes

Yoni Choukroun,Lior Wolf

Error correction code (ECC) is an integral part of the physical communication layer, ensuring reliable data transfer over noisy channels.

Recently, neural decoders have demonstrated their advantage over classical decoding techniques.

However, recent state-of-the-art neural decoders suffer from high complexity and lack the important iterative scheme characteristic of many legacy decoders.

In this work, we propose to employ denoising diffusion models for the soft decoding of linear codes at arbitrary block lengths.

Our framework models the forward channel corruption as a series of diffusion steps that can be reversed iteratively.

Three contributions are made: (i) a diffusion process suitable for the decoding setting is introduced, (ii) the neural diffusion decoder is conditioned on the number of parity errors, which indicates the level of corruption at a given step, (iii) a line search procedure based on the code's syndrome obtains the optimal reverse diffusion step size.

The proposed approach demonstrates the power of diffusion models for ECC and is able to achieve state-of-the-art accuracy, outperforming the other neural decoders by sizable margins, even for a single reverse diffusion step.

****

Progressive Purification for Instance-Dependent Partial Label Learning

Ning Xu,Biao Liu,Jiaqi Lv,Congyu Qiao,Xin Geng

Partial-label learning (PLL) aims to train multi-class classifiers from instances with partial labels (PLs)---a PL for an instance is a set of candidate labels where a fixed but unknown candidate is the true label. In the last few years, the instance-independent generation process of PLs has been extensively studied, on the basis of which many practical and theoretical advances have been made in PLL, while relatively less attention has been paid to the practical setting of instance-dependent PLs, namely, the PL depends not only on the true label but the instance itself. In this paper, we propose a theoretically grounded and practically effective approach called progressive purification (POP) for instance-dependent PLL: in each epoch, POP updates the learning model while purifies each PL by progressively moving out false candidate labels for the next epoch of the model training. Theoretically, we prove that POP enlarges the region where the model is reliable by a promising rate, and eventually approximates the Bayes optimal classifier with mild assumptions; technically, POP is flexible with arbitrary losses and compatible with deep networks, so the previous advanced PLL losses can be embedded in it and the performance is often significantly improved.

****

Meta Knowledge Condensation for Federated Learning

Ping Liu,Xin Yu,Joey Tianyi Zhou

Existing federated learning paradigms usually extensively exchange distributed models, rather than original data, at a central solver to achieve a more powerful model. However, this would incur severe communication burden between a server and multiple clients especially when data distributions are heterogeneous. As a result, current federated learning methods often require plenty of communication rounds in training. Unlike existing paradigms, we introduce an alternative perspective to significantly decrease the federate learning communication cost without leaking original data. In this work, we first present a meta knowledge representation method that extracts meta knowledge from distributed clients. The extracted meta knowledge encodes essential information that can be used to improve the current model. As the training progresses, the contributions of the same training samples to a federated model should also vary. Thus, we introduce a dynamic weight assignment mechanism that enables informative samples to contribute adaptively to the current model update. Then, informative meta knowledge from all active clients is sent to the server for model update. Training model on the combined meta knowledge that is regarded as a condense form of original data can significantly mitigate the heterogeneity issues. Moreover, to further ameliorate data heterogeneity, we also exchange meta knowledge among clients as conditional initialisation for meta knowledge extraction. Extensive experiments demonstrate the effectiveness and efficiency of our proposed method. Remarkably, our method outperforms the state-of-the-art by a large margin (from 74.07% to 92.95%) on MNIST with a restricted communication budget (\textit{i.e.}, 10 rounds).

****

Improved Fully Quantized Training via Rectifying Batch Normalization

Kaixin Xu,Jie Lin,Zhe Wang,Peng Hu,Ziyuan Zhao

Quantization-aware Training (QAT) is able to reduce the training cost by quantizing neural network weights and activations in the forward pass and improve the speed at the inference stage. QAT can be extended to Fully-Quantized Training (FQT), which further accelerates the training by quantizing gradients in the backward pass as back-propagation typically occupies half of the training time. Unfortunately, gradient quantization is challenging as Stochastic Gradient Descent (SGD) based training is sensitive to the precision of the gradient signal. Particularly, the noise introduced by gradient quantization accumulates during backward pass, which causes the exploding gradient problem and results in unstable training and significant accuracy drop. Though Batch Normalization (BatchNorm) is a de-facto resort to stabilize training in regular full-precision scenario, we observe that it fails to prevent the gradient explosion when gradient quantizers are injected in the backward pass. Surprisingly, our theory shows that BatchNorm could amplify the noise accumulation, which in turn hastens the explosion of gradients. A BatchNorm rectification method is derived from our theory to suppress the amplification effect and bridge the performance gap between full-precision training and FQT. Adding this simple rectification loss to baselines generates better results than most prior FQT algorithms on various neural network architectures and datasets, regardless of the gradient bit-widths used (8,4, and 2 bits).

****

Exploring Active 3D Object Detection from a Generalization Perspective

Yadan Luo,Zhuoxiao Chen,Zijian Wang,Xin Yu,Zi Huang,Mahsa Baktashmotlagh

To alleviate the high annotation cost in LiDAR-based 3D object detection, active learning is a promising solution that learns to select only a small portion of unlabeled data to annotate, without compromising model performance. Our empirical study, however, suggests that mainstream uncertainty-based and diversity-based active learning policies are not effective when applied in the 3D detection tas

k, as they fail to balance the trade-off between point cloud informativeness and box-level annotation costs. To overcome this limitation, we jointly investigate three novel criteria in our framework CRB for point cloud acquisition - label conciseness, feature representativeness and
geometric balance, which hierarchically filters out the point clouds of redundant 3D bounding box labels, latent features and geometric characteristics (e.g., point cloud density) from the unlabeled sample pool and greedily selects informative ones with fewer objects to annotate. Our theoretical analysis demonstrates that the proposed criteria aligns the marginal distributions of the selected subset and the prior distributions of the unseen test set, and minimizes the upper bound of the generalization error. To validate the effectiveness and applicability of CRB, we conduct extensive experiments on the two benchmark 3D object detection datasets of KITTI and Waymo and examine both one-stage (i.e., Second) and two-stage 3D detector (i.e., PV-RCNN). Experiments evidence that the proposed approach outperforms existing active learning strategies and achieves fully supervised performance requiring 1% and 8% annotations of bounding boxes and point clouds, respectively.

****

Masked Frequency Modeling for Self-Supervised Visual Pre-Training
Jiahao Xie,Wei Li,Xiaohang Zhan,Ziwei Liu,Yew-Soon Ong,Chen Change Loy
We present Masked Frequency Modeling (MFM), a unified frequency-domain-based approach for self-supervised pre-training of visual models. Instead of randomly inserting mask tokens to the input embeddings in the spatial domain, in this paper, we shift the perspective to the frequency domain. Specifically, MFM first masks out a portion of frequency components of the input image and then predicts the missing frequencies on the frequency spectrum. Our key insight is that predicting masked components in the frequency domain is more ideal to reveal underlying image patterns rather than predicting masked patches in the spatial domain, due to the heavy spatial redundancy. Our findings suggest that with the right configuration of mask-and-predict strategy, both the structural information within high-frequency components and the low-level statistics among low-frequency counterparts are useful in learning good representations. For the first time, MFM demonstrates that, for both ViT and CNN, a simple non-Siamese framework can learn meaningful representations even using none of the following: (i) extra data, (ii) extra model, (iii) mask token. Experimental results on image classification and semantic segmentation, as well as several robustness benchmarks show the competitive performance and advanced robustness of MFM compared with recent masked image modeling approaches. Furthermore, we also comprehensively investigate the effectiveness of classical image restoration tasks for representation learning from a unified frequency perspective and reveal their intriguing relations with our MFM approach. Project page: https://www.mmlab-ntu.com/project/mfm/index.html.

****

Dynamic Prompt Learning via Policy Gradient for Semi-structured Mathematical Reasoning
Pan Lu,Liang Qiu,Kai-Wei Chang,Ying Nian Wu,Song-Chun Zhu,Tanmay Rajpurohit,Peter Clark,Ashwin Kalyan
Mathematical reasoning, a core ability of human intelligence, presents unique challenges for machines in abstract thinking and logical reasoning. Recent large pre-trained language models such as GPT-3 have achieved remarkable progress on mathematical reasoning tasks written in text form, such as math word problems (MWP). However, it is unknown if the models can handle more complex problems that involve math reasoning over heterogeneous information, such as tabular data. To fill the gap, we present Tabular Math Word Problems (TabMWP), a new dataset containing 38,431 open-domain grade-level problems that require mathematical reasoning on both textual and tabular data. Each question in TabMWP is aligned with a tabular context, which is presented as an image, semi-structured text, and a structured table. There are two types of questions: free-text and multi-choice, and each problem is annotated with gold solutions to reveal the multi-step reasoning process. We evaluate different pre-trained models on TabMWP, including the GPT-3 model in a few-shot setting. As earlier studies suggest, since few-shot GPT-3 re

lies on the selection of in-context examples, its performance is unstable and can degrade to near chance. The unstable issue is more severe when handling complex problems like TabMWP. To mitigate this, we further propose a novel approach, PromptPG, which utilizes policy gradient to learn to select in-context examples from a small amount of training data and then constructs the corresponding prompt for the test example. Experimental results show that our method outperforms the best baseline by 5.31% on the accuracy metric and reduces the prediction variance significantly compared to random selection, which verifies its effectiveness in selecting in-context examples. The data and code are available at https://promptpg.github.io.

****

Lottery Aware Sparsity Hunting: Enabling Federated Learning on Resource-Limited Edge

Sara Babakniya,Souvik Kundu,Saurav Prakash,Yue Niu,Salman Avestimehr

Limited computation and communication capabilities of clients pose significant challenges in federated learning (FL) over resource-limited edge nodes. A potential solution to this problem is to deploy off-the-shelf sparse learning algorithms that train a binary sparse mask on each client with the expectation of training a consistent sparse server mask yielding sparse weight tensors. However, as we investigate in this paper, such naive deployments result in a significant drop in accuracy compared to FL with dense models, especially for clients with limited resource budgets. In particular, our investigations reveal a serious lack of consensus among the trained sparsity masks on clients, which prevents convergence for the server mask and potentially leads to a substantial drop in model performance. Based on such key observations, we propose federated lottery aware sparsity hunting (FLASH), a unified sparse learning framework to make the server win a lottery in terms of yielding a sparse sub-model, able to maintain classification performance under highly resource-limited client settings. Moreover, to support FL on different devices requiring different parameter density, we leverage our findings to present hetero-FLASH, where clients can have different target sparsity budgets based on their device resource limits. Experimental evaluations with multiple models on various datasets (both IID and non-IID) show superiority of our models in closing the gap with unpruned baseline while yielding up to ~10.1% improved accuracy with ~10.26x fewer communication costs, compared to existing alternatives, at similar hyperparameter settings.

****

Neuro-Symbolic Procedural Planning with Commonsense Prompting

Yujie Lu,Weixi Feng,Wanrong Zhu,Wenda Xu,Xin Eric Wang,Miguel Eckstein,William Yang Wang

Procedural planning aims to implement complex high-level goals by decomposition into simpler low-level steps. Although procedural planning is a basic skill set for humans in daily life, it remains a challenge for large language models (LLMs) that lack a deep understanding of the cause-effect relations in procedures. Previous methods require manual exemplars to acquire procedural planning knowledge from LLMs in the zero-shot setting. However, such elicited pre-trained knowledge in LLMs induces spurious correlations between goals and steps, which impair the model generalization to unseen tasks. In contrast, this paper proposes a neuro-symbolic procedural PLANner (PLAN) that elicits procedural planning knowledge from the LLMs with commonsense-infused prompting. To mitigate spurious goal-step correlations, we use symbolic program executors on the latent procedural representations to formalize prompts from commonsense knowledge bases as a causal intervention toward the Structural Causal Model. Both automatic and human evaluations on WikiHow and RobotHow show the superiority of PLAN on procedural planning without further training or manual exemplars.

****

Learning Object-Language Alignments for Open-Vocabulary Object Detection

Chuang Lin,Peize Sun,Yi Jiang,Ping Luo,Lizhen Qu,Gholamreza Haffari,Zehuan Yuan,Jianfei Cai

Existing object detection methods are bounded in a fixed-set vocabulary by costly labeled data. When dealing with novel categories, the model has to be retraine d with more bounding box annotations. Natural language supervision is an attractive alternative for its annotation-free attributes and broader object concepts. However, learning open-vocabulary object detection from language is challenging since image-text pairs do not contain fine-grained object-language alignments. Previous solutions rely on either expensive grounding annotations or distilling classification-oriented vision models. In this paper, we propose a novel open-vocabulary object detection framework directly learning from image-text pair data. We formulate object-language alignment as a set matching problem between a set of image region features and a set of word embeddings. It enables us to train an open-vocabulary object detector on image-text pairs in a much simple and effective way. Extensive experiments on two benchmark datasets, COCO and LVIS, demonstrate our superior performance over the competing approaches on novel categories, e.g. achieving 32.0% mAP on COCO and 21.7% mask mAP on LVIS. Code will be released.

****

Phase transition for detecting a small community in a large network

Jiashun Jin,Tracy Ke,Paxton Turner,Anru Zhang

How to detect a small community in a large network is an interesting problem, including clique detection as a special case, where a naive degree-based χ^2-test was shown to be powerful in the presence of an Erdös-Renyi (ER) background. Using Sinkhorn's theorem, we show that the signal captured by the χ^2-test may be a modeling artifact, and it may disappear once we replace the Erdös-Renyi model by a broader network model. We show that the recent SgnQ test is more appropriate for such a setting. The test is optimal in detecting communities with sizes comparable to the whole network, but has never been studied for our setting, which is substantially different and more challenging. Using a degree-corrected block model (DCBM), we establish phase transitions of this testing problem concerning the size of the small community and the edge densities in small and large communities. When the size of the small community is larger than $\sqrt{n}$, the SgnQ test is optimal for it attains the computational lower bound (CLB), the information lower bound for methods allowing polynomial computation time. When the size of the small community is smaller than $\sqrt{n}$, we establish the parameter regime where the SgnQ test has full power and make some conjectures of the CLB. We also study the classical information lower bound (LB) and show that there is always a gap between the CLB and LB in our range of interest. ■

****

On the Word Boundaries of Emergent Languages Based on Harris's Articulation Scheme

Ryo Ueda,Taiga Ishii,Yusuke Miyao

This paper shows that emergent languages in signaling games lack meaningful word boundaries in terms of Harris's Articulation Scheme (HAS), a universal property of natural language. Emergent Languages are artificial communication protocols arising among agents. However, it is not obvious whether such a simulated language would have the same properties as natural language. In this paper, we test if they satisfy HAS. HAS states that word boundaries can be obtained solely from phonemes in natural language. We adopt HAS-based word segmentation and verify whether emergent languages have meaningful word segments. The experiment suggested they do not have, although they meet some preconditions for HAS. We discovered a gap between emergent and natural languages to be bridged, indicating that the standard signaling game satisfies prerequisites but is still missing some necessary ingredients.

****

TempCLR: Temporal Alignment Representation with Contrastive Learning

Yuncong Yang,Jiawei Ma,Shiyuan Huang,Long Chen,Xudong Lin,Guangxing Han,Shih-Fu Chang

Video representation learning has been successful in video-text pre-training for zero-shot transfer, where each sentence is trained to be close to the paired video clips in a common feature space. For long videos, given a paragraph of description where the sentences describe different segments of the video, by matching all sentence-clip pairs, the paragraph and the full video are aligned implicit ly. However, such unit-level similarity measure may ignore the global temporal context over a long time span, which inevitably limits the generalization ability. In this paper, we propose a contrastive learning framework TempCLR to compare the full video and the paragraph explicitly. As the video/paragraph is formulated as a sequence of clips/sentences, under the constraint of their temporal order, we use dynamic time warping to compute the minimum cumulative cost over sentence-clip pairs as the sequence-level distance. To explore the temporal dynamics, we break the consistency of temporal order by shuffling the video clips or sentences according to the temporal granularity. In this way, we obtain the representations for clips/sentences, which perceive the temporal information and thus facilitate the sequence alignment. In addition to pre-training on the video and paragraph, our approach can also generalize on the matching between different video instances. We evaluate our approach on video retrieval, action step localization, and few-shot action recognition, and achieve consistent performance gain over all three tasks. Detailed ablation studies are provided to justify the approach design.

****

Generative Augmented Flow Networks

Ling Pan,Dinghuai Zhang,Aaron Courville,Longbo Huang,Yoshua Bengio

The Generative Flow Network is a probabilistic framework where an agent learns a stochastic policy for object generation, such that the probability of generating an object is proportional to a given reward function. Its effectiveness has been shown in discovering high-quality and diverse solutions, compared to reward-maximizing reinforcement learning-based methods. Nonetheless, GFlowNets only learn from rewards of the terminal states, which can limit its applicability. Indeed, intermediate rewards play a critical role in learning, for example from intrinsic motivation to provide intermediate feedback even in particularly challenging sparse reward tasks. Inspired by this, we propose Generative Augmented Flow Networks (GAFlowNets), a novel learning framework to incorporate intermediate rewards into GFlowNets. We specify intermediate rewards by intrinsic motivation to tackle the exploration problem in sparse reward environments. GAFlowNets can leverage edge-based and state-based intrinsic rewards in a joint way to improve exploration. Based on extensive experiments on the GridWorld task, we demonstrate the effectiveness and efficiency of GAFlowNet in terms of convergence, performance, and diversity of solutions. We further show that GAFlowNet is scalable to a more complex and large-scale molecule generation domain, where it achieves consistent and significant performance improvement.

****

Inferring Fluid Dynamics via Inverse Rendering

Jinxian Liu,Ye Chen,Bingbing Ni,Jiyao Mao,Zhenbo Yu

Humans have a strong intuitive understanding of physical processes such as fluid falling by just a glimpse of such a scene picture, i.e., quickly derived from our immersive visual experiences in memory. This work achieves such a photo-to-fluid-dynamics reconstruction functionality learned from unannotated videos, without any supervision of ground-truth fluid dynamics. In a nutshell, a differentiable Euler simulator modeled with a ConvNet-based pressure projection solver, is integrated with a volumetric renderer, supporting end-to-end/coherent differentiable dynamic simulation and rendering. By endowing each sampled point with a fluid volume value, we derive a NeRF-like differentiable renderer dedicated from fluid data; and thanks to this volume-augmented representation, fluid dynamics could be inversely inferred from error signal between the rendered result and ground-truth video frame (i.e., inverse rendering). Experiments on our generated Fluid Fall datasets and DPI Dam Break dataset are conducted to demonstrate both effectiveness and generalization ability of our method.

****

How Does Value Distribution in Distributional Reinforcement Learning Help Optimization?

Ke Sun,Bei Jiang,Linglong Kong

We consider the problem of learning a set of probability distributions from the Bellman dynamics in distributional reinforcement learning~(RL) that learns the w hole return distribution compared with only its expectation in classical RL. Despite its success to obtain superior performance, we still have a poor understanding of how the value distribution in distributional RL works. In this study, we analyze the optimization benefits of distributional RL by leverage of additional value distribution information over classical RL in the Neural Fitted Z-Iteration~(Neural FZI) framework. To begin with, we demonstrate that the distribution loss of distributional RL has desirable smoothness characteristics and hence enjoys stable gradients, which is in line with its tendency to promote optimization stability. Furthermore, the acceleration effect of distributional RL is revealed by decomposing the return distribution. It turns out that distributional RL can perform favorably if the value distribution approximation is appropriate, measured by the variance of gradient estimates in each environment for any specific distributional RL algorithm. Rigorous experiments validate the stable optimization behaviors of distributional RL, contributing to its acceleration effects compared to classical RL. The findings of our research illuminate how the value distribution in distributional RL algorithms helps the optimization.

****

Bort: Towards Explainable Neural Networks with Bounded Orthogonal Constraint

Borui Zhang,Wenzhao Zheng,Jie Zhou,Jiwen Lu

Deep learning has revolutionized human society, yet the black-box nature of deep neural networks hinders further application to reliability-demanded industries. In the attempt to unpack them, many works observe or impact internal variables to improve the comprehensibility and invertibility of the black-box models. However, existing methods rely on intuitive assumptions and lack mathematical guarantees. To bridge this gap, we introduce Bort, an optimizer for improving model explainability with boundedness and orthogonality constraints on model parameters, derived from the sufficient conditions of model comprehensibility and invertibility. We perform reconstruction and backtracking on the model representations optimized by Bort and observe a clear improvement in model explainability. Based on Bort, we are able to synthesize explainable adversarial samples without additional parameters and training. Surprisingly, we find Bort constantly improves the classification accuracy of various architectures including ResNet and DeiT on MNIST, CIFAR-10, and ImageNet. Code: https://github.com/zbr17/Bort.

****

Interpreting Distributional Reinforcement Learning: A Regularization Perspective

Ke Sun,Yingnan Zhao,Yi Liu,Enze Shi,Yafei Wang,Xiaodong Yan,Bei Jiang,Linglong Kong

Distributional reinforcement learning~(RL) is a class of state-of-the-art algorithms that estimate the entire distribution of the total return rather than its expected value alone. The theoretical advantages of distributional RL over expectation-based RL remain elusive, despite the remarkable performance of distributional RL. Our work attributes the superiority of distributional RL to its regularization effect stemming from the value distribution information regardless of only its expectation. We decompose the value distribution into its expectation and the remaining distribution part using a variant of the gross error model in robust statistics. Hence, distributional RL has an additional benefit over expectation-based RL thanks to the impact of a \textit{risk-sensitive entropy regularization} within the Neural Fitted Z-Iteration framework. Meanwhile, we investigate the role of the resulting regularization in actor-critic algorithms by bridging the risk-sensitive entropy regularization of distributional RL and the vanilla entropy in maximum entropy RL. It reveals that distributional RL induces an augmented reward function, which promotes a risk-sensitive exploration against the intrinsic uncertainty of the environment. Finally, extensive experiments verify the importance of the regularization effect in distributional RL, as well as the mutual impacts of different entropy regularizations. Our study paves the way towards a better understanding of distributional RL, especially when looked at through a regularization lens.

****

The Power of Regularization in Solving Extensive-Form Games

Mingyang Liu,Asuman E. Ozdaglar,Tiancheng Yu,Kaiqing Zhang

In this paper, we investigate the power of {\it regularization}, a common technique in reinforcement learning and optimization, in solving extensive-form games (EFGs). We propose a series of new algorithms based on regularizing the payoff functions of the game, and establish a set of convergence results that strictly improve over the existing ones, with either weaker assumptions or stronger convergence guarantees. In particular, we first show that dilated optimistic mirror descent (DOMD), an efficient variant of OMD for solving EFGs, with adaptive regularization can achieve a fast $\tilde O(1/T)$ last-iterate convergence in terms of duality gap and distance to the set of Nash equilibrium (NE) without uniqueness assumption of the NE. Second, we show that regularized counterfactual regret minimization (\texttt{Reg-CFR}), with a variant of optimistic mirror descent algorithm as regret-minimizer, can achieve $O(1/T^{1/4})$ best-iterate, and $O(1/T^{3/4})$ average-iterate convergence rate for finding NE in EFGs. Finally, we show that \texttt{Reg-CFR} can achieve asymptotic last-iterate convergence, and optimal $O(1/T)$ average-iterate convergence rate, for finding the NE of perturbed EFGs, which is useful for finding approximate extensive-form perfect equilibria (EFPE). To the best of our knowledge, they constitute the first last-iterate convergence results for CFR-type algorithms, while matching the state-of-the-art average-iterate convergence rate in finding NE for non-perturbed EFGs. We also provide numerical results to corroborate the advantages of our algorithms.
****

Neural Topic Modeling with Embedding Clustering Regularization
Xiaobao Wu,Xinshuai Dong,Thong Thanh Nguyen,Anh Tuan Luu

Topic models have been prevalent for decades with various applications like automatic text analysis due to their effectiveness and interpretability. However, existing topic models commonly suffer from the notorious topic collapsing issue: the discovered topics semantically collapse towards each other, leading to highly repetitive topics, insufficient topic discovery, and damaged model interpretability. In this paper, we propose a new neural topic model, Embedding Clustering Regularization Topic Model (ECRTM), to solve the topic collapsing issue. In addition to the reconstruction error of existing work, we propose a novel Embedding Clustering Regularization (ECR), which forces each topic embedding to be the center of a separately aggregated word embedding cluster in the semantic space. Instead of collapsing together, this makes topic embeddings away from each other and cover different semantics of word embeddings. Thus our ECR enables each produced topic to contain distinct word semantics, which alleviates topic collapsing. Through jointly optimizing our ECR objective and the neural topic modeling objective, ECRTM generates diverse and coherent topics together with high-quality topic distributions of documents. Extensive experiments on benchmark datasets demonstrate that ECRTM effectively addresses the topic collapsing issue and consistently surpasses state-of-the-art baselines in terms of topic quality, topic distributions of documents, and downstream classification tasks.

****

Distributional Reinforcement Learning via Sinkhorn Iterations
Ke Sun,Yingnan Zhao,Yi Liu,Wulong Liu,Bei Jiang,Linglong Kong

Distributional reinforcement learning~(RL) is a class of state-of-the-art algorithms that estimate the entire distribution of the total return rather than only its expectation. The empirical success of distributional RL is determined by the representation of return distributions and the choice of distribution divergence. In this paper, we propose a new class of \textit{Sinkhorn distributional RL~(SinkhornDRL)} algorithm that learns a finite set of statistics, i.e., deterministic samples, from each return distribution and then uses Sinkhorn iterations to evaluate the Sinkhorn distance between the current and target Bellmen distributions. Sinkhorn divergence features as the interpolation between the Wasserstein distance and Maximum Mean Discrepancy~(MMD). SinkhornDRL finds a sweet spot by taking advantage of the geometry of optimal transport-based distance and the unbiased gradient estimate property of MMD. Finally, compared to state-of-the-art algorithms, SinkhornDRL's competitive performance is demonstrated on the suit of 55 Atari games.

****

Contextual Symbolic Policy For Meta-Reinforcement Learning

Jiaming Guo,Rui Zhang,Shaohui Peng,Qi Yi,Xing Hu,Ruizhi Chen,Kehan Long,Zidong Du,Xishan Zhang,Ling Li,Qi Guo,Yunji Chen

Context-based Meta-Reinforcement Learning (Meta-RL), which conditions the RL agent on the context variables, is a powerful method for learning a generalizable agent.

Current context-based Meta-RL methods often construct their contextual policy with a neural network (NN) and directly take the context variables as a part of the input. However, the NN-based policy contains tremendous parameters which possibly result in overfitting, the difficulty of deployment and poor interpretability.

To improve the generation ability, efficiency and interpretability, we propose a novel Contextual Symbolic Policy (CSP) framework, which generates contextual policy with a symbolic form based on the context variables for unseen tasks in meta-RL. Our key insight is that the symbolic expression is capable of capturing complex relationships by composing various operators and has a compact form that helps strip out irrelevant information. Thus, the CSP learns to produce symbolic policy for meta-RL tasks and extract the essential common knowledge to achieve higher generalization ability. Besides, the symbolic policies with a compact form are efficient to be deployed and easier to understand.

In the implementation, we construct CSP as a gradient-based framework to learn the symbolic policy from scratch in an end-to-end and differentiable way. The symbolic policy is represented by a symbolic network composed of various symbolic operators. We also employ a path selector to decide the proper symbolic form of the policy and a parameter generator to produce the coefficients of the symbolic policy. Empirically, we evaluate the proposed CSP method on several Meta-RL tasks and demonstrate that the contextual symbolic policy achieves higher performance and efficiency and shows the potential to be interpretable.

****

MLPInit: Embarrassingly Simple GNN Training Acceleration with MLP Initialization

Xiaotian Han,Tong Zhao,Yozen Liu,Xia Hu,Neil Shah

Training graph neural networks (GNNs) on large graphs is complex and extremely time consuming. This is attributed to overheads caused by sparse matrix multiplication, which are sidestepped when training multi-layer perceptrons (MLPs) with only node features. MLPs, by ignoring graph context, are simple and faster for graph data, however they usually sacrifice prediction accuracy, limiting their applications for graph data. We observe that for most message passing-based GNNs, we can trivially derive an analog MLP (we call this a PeerMLP) with an equivalent weight space, by setting the trainable parameters with the same shapes, making us curious about how do GNNs using weights from a fully trained PeerMLP perform? Surprisingly, we find that GNNs initialized with such weights significantly outperform their PeerMLPs, motivating us to use PeerMLP training as a precursor, initialization step to GNN training. To this end, we propose an embarrassingly simple, yet hugely effective initialization method for GNN training acceleration, called \mlpinit. Our extensive experiments on multiple large-scale graph datasets with diverse GNN architectures validate that MLPInit can accelerate the training of GNNs (up to 33× speedup on OGB-Products) and often improve prediction performance (e.g., up to 7.97% improvement for GraphSAGE across 7 datasets for node classification, and up to 17.81% improvement across 4 datasets for link prediction on metric Hits@10). The code is available at https://github.com/snap-research/MLPInit-for-GNNs.

****

Progressively Compressed Auto-Encoder for Self-supervised Representation Learning

Jin Li,Yaoming Wang,XIAOPENG ZHANG,Yabo Chen,Dongsheng Jiang,Wenrui Dai,Chenglin Li,Hongkai Xiong,Qi Tian

As a typical self-supervised learning strategy, Masked Image Modeling (MIM) is driven by recovering all masked patches from visible ones. However, patches from the same image are highly correlated and it is redundant to reconstruct all the masked patches. We find that this redundancy is neglected by existing MIM based methods and causes non-negligible overheads in computation that do not necessarily benefit self-supervised representation. In this paper, we present a novel approach named PCAE, short for Progressively Compressed AutoEncoder, to address the redundant reconstruction issue by progressively compacting tokens and only retaining necessary information for forward propagation and reconstruction. In particular, we identify those redundant tokens in an image via a simple yet effective similarity metric between each token with the mean of the token sequence. Those redundant tokens that other ones can probably represent are progressively dropped accordingly during the forward propagation, and importantly, we only focus on reconstructing these retained tokens. As a result, we are able to achieve a better trade-off between performance and efficiency for pre-training. Besides, benefitting from the flexible strategy, PCAE can be also directly employed for downstream fine-tuning tasks and enable scalable deployment. Experiments show that PCAE achieves comparable performance to MAE with only 1/8 GPU days. The code is available at https://github.com/caddyless/PCAE/.

****

ConBaT: Control Barrier Transformer for Safety-Critical Policy Learning

Yue Meng,Sai Vemprala,Rogerio Bonatti,Chuchu Fan,Ashish Kapoor

Large-scale self-supervised models have recently revolutionized our ability to perform a variety of tasks within the vision and language domains. However, using such models for autonomous systems is challenging because of safety requirements: besides executing correct actions, an autonomous agent needs to also avoid high cost and potentially fatal critical mistakes. Traditionally, self-supervised training mostly focuses on imitating previously observed behaviors, and the training demonstrations carry no notion of which behaviors should be explicitly avoided. In this work, we propose Control Barrier Transformer (ConBaT), an approach that learns safe behaviors from demonstrations in a self-supervised fashion. ConBaT is inspired by the concept of control barrier functions in control theory and uses a causal transformer that learns to predict safe robot actions autoregressively using a critic that requires minimal safety data labeling. During deployment, we employ a lightweight online optimization to find actions that can ensure future states lie within the safe set. We apply our approach to different simulated control tasks and show that our method results in safer control policies compared to other classical and learning-based methods.

****

Robust Transfer Learning Based on Minimax Principle

Xinyi Tong,Xiangxiang Xu,Shao-Lun Huang,Lizhong Zheng

The similarity between target and source tasks is a crucial quantity for theoretical analyses and algorithm designs in transfer learning studies. However, this quantity is often difficult to be precisely captured. To address this issue, we make a boundedness assumption on the task similarity and then propose a mathematical framework based on the minimax principle, which minimizes the worst-case expected population risk under this assumption. Furthermore, our proposed minimax problem can be solved analytically, which provides a guideline for designing robust transfer learning models. According to the analytical expression, we interpret the influences of sample sizes, task distances, and the model dimensionality in knowledge transferring. Then, practical algorithms are developed based on the theoretical results. Finally, experiments conducted on image classification tasks show that our approaches can achieve robust and competitive accuracies under random selections of training sets.

****

Interpreting Neural Networks Through the Lens of Heat Flow

Rui Xia,Fan Yang,Liang Sun

Machine learning models are often developed in a way that prioritizes task-specific performance but defers the understanding of how they actually work. This is especially true nowadays for deep neural networks. In this paper, we step back and consider the basic problem of understanding a learned model represented as a smooth scalar-valued function. We introduce HeatFlow, a framework based upon the heat diffusion process for interpreting the multi-scale behavior of the model a round a test point. At its core, our approach looks into the heat flow initialized at the function of interest, which generates a family of functions with increasing smoothness. By applying differential operators to these smoothed functions, summary statistics (i.e., explanations) characterizing the original model on different scales can be drawn. We place an emphasis on studying the heat flow on data manifold, where the model is trained and expected to be well behaved. Numeric approximation procedures for implementing the proposed method in practice are discussed and demonstrated on image recognition tasks.

****

Efficient Surrogate Gradients for Training Spiking Neural Networks

Hao Lin,Shikuang Deng,Shi Gu

Spiking Neural Network (SNN) is widely regarded as one of the next-generation neural network infrastructures, yet it suffers from an inherent non-differentiable problem that makes the traditional backpropagation (BP) method infeasible. Surrogate gradients (SG), which are an approximation to the shape of the Dirac's δ-function, can help alleviate this issue to some extent. To our knowledge, the majority of research, however, keep a fixed surrogate gradient for all layers, ignorant of the fact that there exists a trade-off between the approximation to the delta function and the effective domain of gradients under the given dataset, hence limiting the efficiency of surrogate gradients and impairing the overall model performance. To guide the shape optimization in applying surrogate gradients for training SNN, we propose an indicator χ, which represents the proportion of parameters with non-zero gradients in backpropagation. Further we present a novel χ-based training pipeline that adaptively makes trade-offs between the surrogate gradients' shapes and its effective domain, followed by a series of ablation experiments for verification. Our algorithm achieves 69.09\% accuracy on the ImageNet dataset using SEW-ResNet34 - a 2.05\% absolute improvement from baseline. Moreover, our method only requires extremely low external cost and can be simply integrated into the existing training procedure.

****

The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning

Zhenmei Shi,Jiefeng Chen,Kunyang Li,Jayaram Raghuram,Xi Wu,Yingyu Liang,Somesh Jha

Pre-training representations (a.k.a. foundation models) has recently become a prevalent learning paradigm, where one first pre-trains a representation using large-scale unlabeled data, and then learns simple predictors on top of the representation using small labeled data from the downstream tasks. There are two key desiderata for the representation: label efficiency (the ability to learn an accurate classifier on top of the representation with a small amount of labeled data) and universality (usefulness across a wide range of downstream tasks). In this paper, we focus on one of the most popular instantiations of this paradigm: contrastive learning with linear probing, i.e., learning a linear predictor on the representation pre-trained by contrastive learning. We show that there exists a trade-off between the two desiderata so that one may not be able to achieve both simultaneously.

Specifically, we provide analysis using a theoretical data model and show that, while more diverse pre-training data result in more diverse features for different tasks (improving universality), it puts less emphasis on task-specific features, giving rise to larger sample complexity for down-stream supervised tasks, and thus worse prediction performance. Guided by this analysis, we propose a contrastive regularization method to improve the trade-off. We validate our analysis and method empirically with systematic experiments using real-world datasets and foundation models.

****

S-NeRF: Neural Radiance Fields for Street Views

Ziyang Xie,Junge Zhang,Wenye Li,Feihu Zhang,Li Zhang

Neural Radiance Fields (NeRFs) aim to synthesize novel views of objects and scenes, given the object-centric camera views with large overlaps. However, we conjugate that this paradigm does not fit the nature of the street views that are col lected by many self-driving cars from the large-scale unbounded scenes. Also, the onboard cameras perceive scenes without much overlapping. Thus, existing NeRFs often produce blurs, "floaters" and other artifacts on street-view synthesis. In this paper, we propose a new street-view NeRF (S-NeRF) that considers novel view synthesis of both the large-scale background scenes and the foreground moving vehicles jointly. Specifically, we improve the scene parameterization function and the camera poses for learning better neural representations from street views. We also use the the noisy and sparse LiDAR points to boost the training and learn a robust geometry and reprojection based confidence to address the depth outliers.   Moreover, we extend our S-NeRF for reconstructing moving vehicles that is impracticable for conventional NeRFs. Thorough experiments on the large-scale driving datasets (e.g., nuScenes and Waymo) demonstrate that our method beats the state-of-the-art rivals by reducing 7■40% of the mean-squared error in the street-view synthesis and a 45% PSNR gain for the moving vehicles rendering.

****

Learning Visual Representation with Synthetic Images and Topologically-defined Labels

Shizuo Kaji,Yohsuke Watanabe

We propose a scheme for neural networks to learn visual representation with synthetic images and mathematically-defined labels that capture topological information. To verify that the model acquires a different visual representation than with the usual supervised learning with manually-defined labels, we show that the models pretrained with our scheme can be finetuned for image classification tasks to achieve an improved convergence compared to those trained from scratch.

Convolutional neural networks, built upon iterative local operations, are good at learning local features of the image, such as texture, whereas they tend to pay less attention to larger structures. Our method provides a simple way to encourage the model to learn global features through a specifically designed task based on topology. Furthermore, our method requires no real images nor manual labels; hence it sheds light on some of the lately concerned topics in computer vision, such as the cost and the fairness in data collection and annotation.

****

Cycle-consistent Masked AutoEncoder for Unsupervised Domain Generalization

Haiyang Yang,Xiaotong Li,SHIXIANG TANG,Feng Zhu,Yizhou Wang,Meilin Chen,LEI BAI,Rui Zhao,Wanli Ouyang

Self-supervised learning methods undergo undesirable performance drops when there exists a significant domain gap between training and testing scenarios. Therefore, unsupervised domain generalization (UDG) is proposed to tackle the problem, which requires the model to be trained on several different domains without supervision and generalize well on unseen test domains. Existing methods either rely on a cross-domain and semantically consistent image pair in contrastive methods or the reconstruction pair in generative methods, while the precious image pairs are not available without semantic labels. In this paper, we propose a cycle cross-domain reconstruction task for unsupervised domain generalization in the absence of paired images. The cycle cross-domain reconstruction task converts a masked image from one domain to another domain and then reconstructs the original image from the converted images. To preserve the divergent domain knowledge of decoders in the cycle reconstruction task, we propose a novel domain-contrastive loss to regularize the domain information in reconstructed images encoded with the desirable domain style. Qualitative results on extensive datasets illustrate our method improves the state-of-the-art unsupervised domain generalization methods by average $\textbf{+5.59\%}, \textbf{+4.52\%}, \textbf{+4.22\%}, \textbf{+7.02\%}$ on $1\%, 5\%, 10\%, 100\%$ PACS, and $\textbf{+5.08\%}, \textbf{+6.49\%}, \textbf{+1.79\%}, \textbf{+0.53\%}$ on $1\%, 5\%, 10\%, 100\%$ DomainNet, respectively.

****

CFlowNets: Continuous Control with Generative Flow Networks

Yinchuan Li,Shuang Luo,Haozhi Wang,Jianye HAO

Generative flow networks (GFlowNets), as an emerging technique, can be used as an alternative to reinforcement learning for exploratory control tasks. GFlowNets aims to sample actions with a probability proportional to the reward, similar to sampling different candidates in an active learning fashion. However, existing GFlowNets cannot adapt to continuous control tasks because GFlowNets need to form a DAG and compute the flow matching loss by traversing the inflows and outflows of each node in the trajectory. In this paper, we propose generative continuous flow networks (CFlowNets) that can be applied to continuous control tasks. First, we present the theoretical formulation of CFlowNets. Then, a training framework for CFlowNets is proposed, including the action selection process, the flow approximation algorithm, and the continuous flow matching loss function. Afterward, we theoretically prove the error bound of the flow approximation. The error decreases rapidly as the number of flow samples increases. Finally, experimental results on continuous control tasks demonstrate the performance advantages of CFlowNets compared to many reinforcement learning methods, especially regarding exploration ability.

****

Differentiable Gaussianization Layers for Inverse Problems Regularized by Deep Generative Models

Dongzhuo Li

Deep generative models such as GANs, normalizing flows, and diffusion models are powerful regularizers for inverse problems. They exhibit great potential for helping reduce ill-posedness and attain high-quality results. However, the latent tensors of such deep generative models can fall out of the desired high-dimensional standard Gaussian distribution during inversion, particularly in the presence of data noise and inaccurate forward models, leading to low-fidelity solutions. To address this issue, we propose to reparameterize and Gaussianize the latent tensors using novel differentiable data-dependent layers wherein custom operators are defined by solving optimization problems. These proposed layers constrain inverse problems to obtain high-fidelity in-distribution solutions. We validate our technique on three inversion tasks: compressive-sensing MRI, image deblurring, and eikonal tomography (a nonlinear PDE-constrained inverse problem) using two representative deep generative models: StyleGAN2 and Glow. Our approach achieves state-of-the-art performance in terms of accuracy and consistency.

****

DBQ-SSD: Dynamic Ball Query for Efficient 3D Object Detection

Jinrong Yang,Lin Song,Songtao Liu,Weixin Mao,Zeming Li,Xiaoping Li,Hongbin Sun,Jian Sun,Nanning Zheng

Many point-based 3D detectors adopt point-feature sampling strategies to drop some points for efficient inference. These strategies are typically based on fixed and handcrafted rules, making it difficult to handle complicated scenes. Different from them, we propose a Dynamic Ball Query (DBQ) network to adaptively select a subset of input points according to the input features, and assign the feature transform with a suitable receptive field for each selected point. It can be embedded into some state-of-the-art 3D detectors and trained in an end-to-end manner, which significantly reduces the computational cost. Extensive experiments demonstrate that our method can reduce latency by 30%-100% on KITTI, Waymo, and ONCE datasets. Specifically, the inference speed of our detector can reach 162 FPS on KITTI scene, and 30 FPS on Waymo and ONCE scenes without performance degradation. Due to skipping the redundant points, some evaluation metrics show significant improvements.

****

Exploring Low-Rank Property in Multiple Instance Learning for Whole Slide Image Classification

Jinxi Xiang,Jun Zhang

The classification of gigapixel-sized whole slide images (WSIs) with slide-level labels can be formulated as a multiple-instance-learning (MIL) problem. State-of-the-art models often consist of two decoupled parts: local feature embedding with a pre-trained model followed by a global feature aggregation network for classification. We leverage the properties of the apparent similarity in high-resolution WSIs, which essentially exhibit \textit{low-rank} structures in the data manifold, to develop a novel MIL with a boost in both feature embedding and featu re aggregation. We extend the contrastive learning with a pathology-specific Low-Rank Constraint (LRC) for feature embedding to pull together samples (i.e., patches) belonging to the same pathological tissue in the low-rank subspace and simultaneously push apart those from different latent subspaces. At the feature aggregation stage, we introduce an iterative low-rank attention MIL (ILRA-MIL) model to aggregate features with low-rank learnable latent vectors to model global interactions among all instances. We highlight the importance of instance correlation modeling but refrain from directly using the transformer encoder considering the $O(n^2)$ complexity. ILRA-MIL with LRC pre-trained features achieves strong empirical results across various benchmarks, including (i) 96.49\% AUC on the CAMELYON16 for binary metastasis classification, (ii) 97.63\% AUC on the TCGA-NSCLC for lung cancer subtyping, and (iii) 0.6562 kappa on the large-scale PANDA dataset for prostate cancer classification. The code is available at https://github.com/jinxixiang/low_rank_wsi.

****

Evaluating and Inducing Personality in Pre-trained Language Models

Guangyuan Jiang,Manjie Xu,Song-Chun Zhu,Wenjuan Han,Chi Zhang,Yixin Zhu

Originated as a philosophical quest, personality discerns how individuals differ from each other in terms of thinking, feeling, and behaving. Toward building social machines that work with humans on a daily basis, we are motivated to ask: (1) Do existing Large Language Models (LLMs) possess personalities, akin to their human counterparts? (2) If so, how can we evaluate them? (3) Further, given this evaluation framework, how can we induce a certain personality in a fully controllable fashion? To tackle these three questions, we propose the Machine Personality Inventory (MPI) dataset for evaluating the machine personality; MPI follows standardized personality tests, built upon the Big Five Personality Factors (Big Five) theory and personality assessment inventories. By evaluating models with MPI, we provide the first piece of evidence showing the existence of personality in LLMs. We further devise a Chain Prompting method to induce LLMs with a specific personality in a controllable manner, capable of producing diversified behaviors. We hope to shed light on future studies by adopting personality as the essential guide for various downstream tasks, building more human-like and in situ dialogue agents.

****

Node Classification Beyond Homophily: Towards a General Solution

Zhe Xu,Yuzhong Chen,Qinghai Zhou,Yuhang Wu,Menghai Pan,Hao Yang,Hanghang Tong

Graph neural networks (GNNs) have become core building blocks behind a myriad of graph learning tasks. The vast majority of the existing GNNs are built upon, either implicitly or explicitly, the homophily assumption, which is not always true and could heavily degrade the performance of learning tasks. In response, GNNs tailored for heterophilic graphs have been developed. However, most of the existing works are designed for the specific GNN models to address heterophily, which lacks generality. In this paper, we study the problem from the structure learning perspective and propose a family of general solutions named ALT. It can work hand in hand with most of the existing GNNs to decently handle graphs with either low or high homophily. The core of our method is learning to (1) decompose a given graph into two components, (2) extract complementary graph signals from these two components, and (3) adaptively merge the graph signals for node classification. Moreover, analysis based on graph signal processing shows that our framework can empower a broad range of existing GNNs to have adaptive filter characteristics and further modulate the input graph signals, which is critical for handling complex homophilic/heterophilic patterns. The proposed ALT brings significant and consistent performance improvement in node classification for a wide range of GNNs over a variety of real-world datasets.

****

Causal Balancing for Domain Generalization

Xinyi Wang,Michael Saxon,Jiachen Li,Hongyang Zhang,Kun Zhang,William Yang Wang

While machine learning models rapidly advance the state-of-the-art on various real-world tasks, out-of-domain (OOD) generalization remains a challenging problem given the vulnerability of these models to spurious correlations. We propose a balanced mini-batch sampling strategy to transform a biased data distribution into a spurious-free balanced distribution, based on the invariance of the underlying causal mechanisms for the data generation process. We argue that the Bayes optimal classifiers trained on such balanced distribution are minimax optimal across a diverse enough environment space. We also provide an identifiability guarantee of the latent variable model of the proposed data generation process, when utilizing enough train environments. Experiments are conducted on DomainBed, demonstrating empirically that our method obtains the best performance across 20 baselines reported on the benchmark.

****

Neural Radiance Fields with Geometric Consistency for Few-Shot Novel View Synthesis

Jiuhn Song,Min-Seop Kwak,Seungryong Kim

We present a novel method to regularizes neural radiance field (NeRF) in few-shot setting with geometry-based consistency regularization. The proposed approach leverages NeRF's rendered depth map to warp source images to unobserved viewpoints and impose them as pseudo ground truths to facilitate learning of detailed features. By encouraging consistency at feature-level instead of using pixel-level reconstruction loss, we regularize the network solely at semantic and structural levels while allowing view-dependent radiance to model freely after color variations. Our application of proposed consistency term for the network is twofold: between and observed and unobserved viewpoints, image rendered at unseen view is forced to model after the image warped from input observation, while between observed viewpoints the warped image undergoes optimization for geometry-specific regularization. We also demonstrate an effective method to filter out erroneous warped solutions, along with relevant techniques to stabilize training during optimization. We show that our model achieves competitive results compared to concurrent few-shot NeRF models.

****

Towards Addressing Label Skews in One-Shot Federated Learning

Yiqun Diao,Qinbin Li,Bingsheng He

Federated learning (FL) has been a popular research area, where multiple clients collaboratively train a model without sharing their local raw data. Among existing FL solutions, one-shot FL is a promising and challenging direction, where the clients conduct FL training with a single communication round. However, while label skew is a common real-world scenario where some clients may have few or no data of some classes, existing one-shot FL approaches that conduct voting on the local models are not able to produce effective global models. Due to the limited number of classes in each party, the local models misclassify the data from unseen classes into seen classes, which leads to very ineffective global models from voting. To address the label skew issue in one-shot FL, we propose a novel approach named FedOV which generates diverse outliers and introduces them as an additional unknown class in local training to improve the voting performance. Specifically, based on open-set recognition, we propose novel outlier generation approaches by corrupting the original features and further develop adversarial learning to enhance the outliers. Our extensive experiments show that FedOV can significantly improve the test accuracy compared to state-of-the-art approaches in various label skew settings.

****

Breaking Correlation Shift via Conditional Invariant Regularizer

Mingyang Yi,Ruoyu Wang,Jiacheng Sun,Zhenguo Li,Zhi-Ming Ma

Recently, generalization on out-of-distribution (OOD) data with correlation shift has attracted great attentions. The correlation shift is caused by the spurious attributes that correlate to the class label, as the correlation between them may vary in training and test data. For such a problem, we show that given the class label, the models that are conditionally independent of spurious attributes are OOD generalizable. Based on this, a metric Conditional Spurious Variation (CSV) which controls the OOD generalization error, is proposed to measure such conditional independence. To improve the OOD generalization, we regularize the training process with the proposed CSV. Under mild assumptions, our training object ive can be formulated as a nonconvex-concave mini-max problem. An algorithm with a provable convergence rate is proposed to solve the problem. Extensive empirical results verify our algorithm's efficacy in improving OOD generalization.

****

CROM: Continuous Reduced-Order Modeling of PDEs Using Implicit Neural Representations

Peter Yichen Chen,Jinxu Xiang,Dong Heon Cho,Yue Chang,G A Pershing,Henrique Teles Maia,Maurizio M Chiaramonte,Kevin Thomas Carlberg,Eitan Grinspun

The long runtime of high-fidelity partial differential equation (PDE) solvers makes them unsuitable for time-critical applications. We propose to accelerate PDE solvers using reduced-order modeling (ROM). Whereas prior ROM approaches reduce the dimensionality of discretized vector fields, our continuous reduced-order modeling (CROM) approach builds a low-dimensional embedding of the continuous vector fields themselves, not their discretization. We represent this reduced manifold using continuously differentiable neural fields, which may train on any and all available numerical solutions of the continuous system, even when they are obtained using diverse methods or discretizations. We validate our approach on an extensive range of PDEs with training data from voxel grids, meshes, and point clouds. Compared to prior discretization-dependent ROM methods, such as linear subspace proper orthogonal decomposition (POD) and nonlinear manifold neural-network-based autoencoders, CROM features higher accuracy, lower memory consumption, dynamically adaptive resolutions, and applicability to any discretization. For equal latent space dimension, CROM exhibits 79$\times$ and 49$\times$ better accuracy, and 39$\times$ and 132$\times$ smaller memory footprint, than POD and autoencoder methods, respectively. Experiments demonstrate 109$\times$ and 89$\times$ wall-clock speedups over unreduced models on CPUs and GPUs, respectively. Videos and codes are available on the project page: https://crom-pde.github.io

****

Towards One-shot Neural Combinatorial Solvers: Theoretical and Empirical Notes on the Cardinality-Constrained Case

Runzhong Wang,Li Shen,Yiting Chen,Xiaokang Yang,Dacheng Tao,Junchi Yan

One-shot non-autoregressive neural networks, different from RL-based ones, have been actively adopted for solving combinatorial optimization (CO) problems, which can be trained by the objective score in a self-supervised manner. Such methods have shown their superiority in efficiency (e.g. by parallelization) and potential for tackling predictive CO problems for decision-making under uncertainty. While the discrete constraints often become a bottleneck for gradient-based neural solvers, as currently handled in three typical ways: 1) adding a soft penalty in the objective, where a bounded violation of the constraints cannot be guaranteed, being critical to many constraint-sensitive scenarios; 2) perturbing the input to generate an approximate gradient in a black-box manner, though the constraints are exactly obeyed while the approximate gradients can hurt the performance on the objective score; 3) a compromise by developing soft algorithms whereby the output of neural networks obeys a relaxed constraint, and there can still occur an arbitrary degree of constraint-violation. Towards the ultimate goal of establishing a general framework for neural CO solver with the ability to control an arbitrary-small degree of constraint violation, in this paper, we focus on a more achievable and common setting: the cardinality constraints, which in fact can be readily encoded by a differentiable optimal transport (OT) layer. Based on this observation, we propose OT-based cardinality constraint encoding for end-to-end CO problem learning with two variants: Sinkhorn and Gumbel-Sinkhorn, whereby their violation of the constraints can be exactly characterized and bounded by our theoretical results. On synthetic and real-world CO problem instances, our methods surpass the state-of-the-art CO network and are comparable to (if not superior to) the commercial solver Gurobi. In particular, we further showcase a case study of applying our approach to the predictive portfolio optimization task on real-world asset price data, improving the Sharpe ratio from 1.1 to 2.0 of a strong LSTM+Gurobi baseline under the classic predict-then-optimize paradigm.

****

Block and Subword-Scaling Floating-Point (BSFP) : An Efficient Non-Uniform Quant ization For Low Precision Inference

Yun-Chen Lo,Tse-Kuang Lee,Ren-Shuo Liu

In this paper, we propose Block and Subword-Scaling Floating-Point (BSFP), a non-uniform quantization scheme for the skewed and non-uniform distribution of weight vectors in neural networks. By quantizing each weight vector as the superposition of multiple subword vectors (in two's complement) with scaling factors (in Low-bit Floating-Point, LBFP), BSFP can effectively fit the distribution of weight vectors while maintaining high computation efficiency. Furthermore, we present a grid search-based MSE-optimal quantization flow and a scaled serial processing engine to complete the quantization pipeline and the infrastructure.

The experimental results on the ImageNet classification task show that our proposed method outperforms state-of-the-art Microsoft Floating Point (MSFP) by up to 20.56% top-1 accuracy at the same weight precision and reduces up to 10.3% model size. Furthermore, BSFP outperforms MSFP by up to 2.0$\times$ computing throughput and up to 5.3$\times$ energy efficiency under the same silicon area budget.

****

Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning

Rundong Luo,Yifei Wang,Yisen Wang

Recent works have shown that self-supervised learning can achieve remarkable robustness when integrated with adversarial training (AT). However, the robustness gap between supervised AT (sup-AT) and self-supervised AT (self-AT) remains significant. Motivated by this observation, we revisit existing self-AT methods and discover an inherent dilemma that affects self-AT robustness: either strong or weak data augmentations are harmful to self-AT, and a medium strength is insufficient to bridge the gap. To resolve this dilemma, we propose a simple remedy named DYNACL (Dynamic Adversarial Contrastive Learning). In particular, we propose an augmentation schedule that gradually anneals from a strong augmentation to a weak one to benefit from both extreme cases. Besides, we adopt a fast post-processing stage for adapting it to downstream tasks. Through extensive experiments, we show that DYNACL can improve state-of-the-art self-AT robustness by 8.84% under Auto-Attack on the CIFAR-10 dataset, and can even outperform vanilla supervised adversarial training for the first time. Our code is available at \url{https://github.com/PKU-ML/DYNACL}.

****

Semi-supervised Community Detection via Structural Similarity Metrics

Yicong Jiang,Tracy Ke

Motivated by the interests of social network analysis and network-based recommendation systems, we consider a semi-supervised community detection problem, where the goal is to estimate the community label of a new node by leveraging on the network structure and partially observed community labels of existing nodes.

We model the network with a degree-corrected stochastic block model, which allows for severe degree heterogeneity and potentially non-assortative communities.

We propose a fast algorithm that computes a `structural similarity metric' between the new node and each of the K communities, aggregating information in labeled and unlabeled data. The estimated label of the new node is equal to the value of k that maximizes this similarity metric. Our method is computationally fast and compares favorably with existing semi-supervised algorithms on numerical performance. In theory, we derive explicit bounds for the misclassification error and show the efficiency of our method by comparing it with an ideal classifier. To our best knowledge, our results provide the first semi-supervised community detection algorithm with theoretical guarantees.

****

DDM^2: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models

Tiange Xiang,Mahmut Yurt,Ali B Syed,Kawin Setsompop,Akshay Chaudhari

Magnetic resonance imaging (MRI) is a common and life-saving medical imaging technique. However, acquiring high signal-to-noise ratio MRI scans requires long scan times, resulting in increased costs and patient discomfort, and decreased throughput. Thus, there is great interest in denoising MRI scans, especially for th e subtype of diffusion MRI scans that are severely SNR-limited. While most prior MRI denoising methods are supervised in nature, acquiring supervised training datasets for the multitude of anatomies, MRI scanners, and scan parameters proves impractical. Here, we propose Denoising Diffusion Models for Denoising Diffusion MRI (DDM^2), a self-supervised denoising method for MRI denoising using diffusion denoising generative models. Our three-stage framework integrates statistic-based denoising theory into diffusion models and performs denoising through conditional generation. During inference, we represent input noisy measurements as a sample from an intermediate posterior distribution within the diffusion Markov chain. We conduct experiments on 4 real-world in-vivo diffusion MRI datasets and show that our DDM^2 demonstrates superior denoising performances ascertained with clinically-relevant visual qualitative and quantitative metrics.

****

Multivariate Time-series Imputation with Disentangled Temporal Representations

SHUAI LIU,Xiucheng Li,Gao Cong,Yile Chen,YUE JIANG

Multivariate time series often faces the problem of missing value. Many time series imputation methods have been developed in the literature. However, these methods all rely on an entangled representation to model dynamics of time series, which may fail to fully exploit the multiple factors (e.g., periodic patterns) contained in the time series. Moreover, the entangled representation usually has no semantic meaning, and thus they often lack interpretability. In addition, many recent models are proposed to deal with the whole time series to capture cross-channel correlations and identify temporal dynamics, but they are not scalable to large-scale datasets. Different from existing approaches, we propose TIDER, a novel matrix factorization-based method with disentangled temporal representations that account for multiple factors, namely trend, seasonality, and local bias, to model complex dynamics. The learned disentanglement makes the imputation process more reliable and offers explainability for imputation results. Moreover, TIDER is scalable to large datasets. Empirical results show that our method not only outperforms existing approaches by notable margins on three real-world datasets, but also scales well to large datasets on which existing deep learning based methods struggle. Disentanglement validation experiments further demonstrate the robustness of our model in obtaining accurate and explainable disentangled components.

****

Knowledge-driven Scene Priors for Semantic Audio-Visual Embodied Navigation

Gyan Tatiya,Jonathan Francis,Luca Bondi,Ingrid Navarro,Eric Nyberg,Jivko Sinapov,Jean Oh

Generalisation to unseen contexts remains a challenge for embodied navigation agents. In the context of semantic audio-visual navigation (SAVi) tasks, generalisation includes both generalising to unseen indoor visual scenes as well as generalising to unheard sounding objects. Previous SAVi task definitions do not include evaluation conditions on truly novel sounding objects, resorting instead to evaluating agents on unheard sound clips of known objects; meanwhile, previous SAVi methods do not include explicit mechanisms for incorporating domain knowledge about object and region semantics. These weaknesses limit the development and assessment of models' abilities to generalise their learned experience. In this work, we introduce the use of knowledge-driven scene priors in the semantic audio-visual embodied navigation task: we combine semantic information from our novel knowledge graph that encodes object-region relations, spatial knowledge from dual Graph Encoder Networks, and background knowledge from a series of pre-training tasks---all within a reinforcement learning framework for audio-visual navigation. We define a new audio-visual navigation sub-task, where agents are evaluated on novel sounding objects, as opposed to unheard clips of known objects. We show improvements over strong baselines in generalisation to unseen regions and novel sounding objects, within the Habitat-Matterport3D simulation environment, under the SoundSpaces task. We release code, knowledge graphs, and dataset generation details in the supplementary material.

****

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language

Andy Zeng,Maria Attarian,brian ichter,Krzysztof Marcin Choromanski,Adrian Wong,Stefan Welker,Federico Tombari,Aveek Purohit,Michael S Ryoo,Vikas Sindhwani,Johnny Lee,Vincent Vanhoucke,Pete Florence

We investigate how multimodal prompt engineering can use language as the intermediate representation to combine complementary knowledge from different pretrained (potentially multimodal) language models for a variety of tasks. This approach is both distinct from and complementary to the dominant paradigm of joint multimodal training. It also recalls a traditional systems-building view as in classical NLP pipelines, but with prompting large pretrained multimodal models. We refer to these as Socratic Models (SMs): a modular class of systems in which multiple pretrained models may be composed zero-shot via multimodal-informed prompting to capture new multimodal capabilities, without additional finetuning. We show that these systems provide competitive state-of-the-art performance for zero-shot image captioning and video-to-text retrieval, and also enable new applications such as (i) answering free-form questions about egocentric video, (ii) engaging in multimodal assistive dialogue with people (e.g., for cooking recipes), and (iii) robot perception and planning. We hope this work provides (a) results for stronger zero-shot baseline performance with analysis also highlighting their limitations, (b) new perspectives for building multimodal systems powered by large pretrained models, and (c) practical application advantages in certain regimes limited by data scarcity, training compute, or model access.

****

CAST: Concurrent Recognition and Segmentation with Adaptive Segment Tokens

Tsung-Wei Ke,Jyh-Jing Hwang,Stella Yu

Recognizing an image and segmenting it into coherent regions are often treated as separate tasks. Human vision, however, has a general sense of segmentation hierarchy before recognition occurs. We are thus inspired to learn image recognition with hierarchical image segmentation based entirely on unlabeled images. Our insight is to learn fine-to-coarse features concurrently at superpixels, segments, and full image levels, enforcing consistency and goodness of feature induced segmentations while maximizing discrimination among image instances.

Our model innovates vision transformers on three aspects. 1) We use adaptive segment tokens instead of fixed-shape patch tokens. 2) We create a token hierarchy by inserting graph pooling between transformer blocks, naturally producing consistent multi-scale segmentations while increasing the segment size and reducing the number of tokens. 3) We produce hierarchical image segmentation for free {\it while} training for recognition by maximizing image-wise discrimination.

Our work delivers the first concurrent recognition and hierarchical segmentation model without any supervision.   Validated on ImageNet and PASCAL VOC, it achieves better recognition and segmentation with higher computational efficiency.

****

Multi-lingual Evaluation of Code Generation Models

Ben Athiwaratkun,Sanjay Krishna Gouda,Zijian Wang,Xiaopeng Li,Yuchen Tian,Ming Tan,Wasi Uddin Ahmad,Shiqi Wang,Qing Sun,Mingyue Shang,Sujan Kumar Gonugondla,Hantian Ding,Varun Kumar,Nathan Fulton,Arash Farahani,Siddhartha Jain,Robert Giaquinto,Haifeng Qian,Murali Krishna Ramanathan,Ramesh Nallapati,Baishakhi Ray,Parminder Bhatia,Sudipta Sengupta,Dan Roth,Bing Xiang

We present two new benchmarks, MBXP and Multilingual HumanEval, designed to evaluate code completion models in over 10 programming languages. These datasets are generated using a conversion framework that transpiles prompts and test cases from the original MBPP and HumanEval datasets into the corresponding data in the target language. By using these benchmarks, we are able to assess the performance of code generation models in a multi-lingual fashion, and discovered generalization ability of language models on out-of-domain languages, advantages of multi-lingual models over mono-lingual, the ability of few-shot prompting to teach the model new languages, and zero-shot translation abilities. In addition, we use our code generation model to perform large-scale bootstrapping to obtain synthetic canonical solutions in several languages, which can be used for other code-r elated evaluations such as code insertion, robustness, or summarization tasks.

****

GRACE-C: Generalized Rate Agnostic Causal Estimation via Constraints

Mohammadsajad Abavisani,David Danks,Sergey Plis

Graphical structures estimated by causal learning algorithms from time series data can provide highly misleading causal information if the causal timescale of the generating process fails to match the measurement timescale of the data. Existing algorithms provide limited resources to respond to this challenge, and so researchers must either use models that they know are likely misleading, or else forego causal learning entirely. Existing methods face up-to-four distinct shortfalls, as they might a) require that the difference between causal and measurement timescales is known; b) only handle very small number of random variables when the timescale difference is unknown; c) only apply to pairs of variables (albeit with fewer assumptions about prior knowledge); or d) be unable to find a solution given statistical noise in the data. This paper aims to address these challenges. We present an approach that combines constraint programming with both theoretical insights into the problem structure and prior information about admissible causal interactions to achieve speed up of multiple orders of magnitude. The resulting system scales to significantly larger sets of random variables (>100) without knowledge of the timescale difference while maintaining theoretical guarantees. This method is also robust to edge misidentification and can use parametric connection strengths, while optionally finding the optimal among many possible solutions.

****

How Powerful is Implicit Denoising in Graph Neural Networks

Songtao Liu,Zhitao Ying,Hanze Dong,Lu Lin,Jinghui Chen,Dinghao Wu

Graph Neural Networks (GNNs), which aggregate features from neighbors, are widely used for processing graph-structured data due to their powerful representation learning capabilities. It is generally believed that GNNs can implicitly remove feature noises. However, existing works have not rigorously analyzed the implicit denoising effect in graph neural networks. In this work, we conduct a comprehensive theoretical study and analyze when and why implicit denoising happens in GNNs. Our theoretical analysis suggests that the implicit denoising largely depends on the connectivity and size of the graph, as well as the GNN architectures. Motivated by adversarial machine learning in improving the robustness of neural networks, we propose the adversarial graph signal denoising (AGSD) problem. By solving such a problem, we derive a robust graph convolution, where the smoothness of the node representations and the implicit denoising effect can be enhanced. Extensive empirical evaluations verify our theoretical analyses and the effectiveness of our proposed model.

****

Unified Detoxifying and Debiasing in Language Generation via Inference-time Adaptive Optimization

Zonghan Yang,Xiaoyuan Yi,Peng Li,Yang Liu,Xing Xie

Recently pre-trained language models (PLMs) have prospered in various natural language generation (NLG) tasks due to their ability to generate fairly fluent text. Nevertheless, these models are observed to capture and reproduce harmful contents in training corpora, typically toxic language and social biases, raising severe moral issues. Prior works on ethical NLG tackle detoxifying and debiasing separately, which is problematic since we find debiased models still exhibit toxicity while detoxified ones even exacerbate biases. To address such a challenge, we propose the first unified framework of detoxifying and debiasing called UDDIA, which jointly formalizes these two problems as rectifying the output space. We theoretically interpret our framework as learning a text distribution mixing weighted attributes. Besides, UDDIA conducts adaptive optimization of only a few parameters during decoding based on a parameter-efficient tuning schema without any training data. This leads to minimal generation quality loss and improved rectification performance with acceptable computational cost. Experimental results demonstrate that compared to several strong baselines, UDDIA achieves debiasing and detoxifying simultaneously and better balances efficiency and effectiveness, taking a further step towards practical ethical NLG.

****

Distribution Aware Metrics for Conditional Natural Language Generation

David Chan,Yiming Ni,Sudheendra Vijayanarasimhan,David A Ross,Austin Myers,John Canny

Traditional automated metrics for evaluating conditional natural language generation use pairwise comparisons between a single generated text and the best-matching gold-standard ground truth text. When multiple ground truths are available, scores are aggregated using an average or max operation across references. While this approach works well when diversity in the ground truth data (i.e. dispersion of the distribution of conditional texts) can be ascribed to noise, such as in automated speech recognition, it does not allow for robust evaluation in the case where diversity in the ground truths represents signal for the model. In this work we argue that existing metrics are not appropriate for domains such as visual description or summarization where ground truths are semantically diverse, and where the diversity in those captions captures useful additional information about the context. We propose a novel paradigm for multi-candidate evaluation of conditional language generation models, and a new family of metrics that compare the {\em distributions} of reference and model-generated caption sets using small sample sets of each. We demonstrate the utility of our approach with a case study in visual description: where we show that existing models optimize for single-description quality over diversity, and gain some insights into how sampling methods and temperature impact description quality and diversity.

****

Recommender Transformers with Behavior Pathways

Zhiyu Yao,Xinyang Chen,Sinan Wang,Qinyan Dai,Yumeng Li,Tanchao Zhu,Mingsheng Long

Sequential recommendation requires the recommender to capture the evolving behavior characteristics from logged user behavior data for accurate recommendations. Nevertheless, user behavior sequences are viewed as a script with multiple ongoing threads intertwined. We find that only a small set of pivotal behaviors can be evolved into the user's future action. As a result, the future behavior of the user is hard to predict. We conclude this characteristic for sequential behaviors of each user as the \textit{behavior pathway}. Different users have their unique behavior pathways. Among existing sequential models, transformers have shown great capacity in capturing global-dependent characteristics. However, these models mainly provide a dense distribution over all previous behaviors using the self-attention mechanism, making the final predictions overwhelmed by the trivial behaviors not adjusted to each user. In this paper, we build the Recommender Transformer (RETR) with a novel Pathway Attention mechanism. RETR can dynamically plan the behavior pathway specified for each user, and sparingly activate the network through this behavior pathway to effectively capture evolving patterns useful for recommendation. The key design is a learned binary route to prevent the behavior pathway from being overwhelmed by trivial behaviors. Pathway attention is model-agnostic and can be applied to a series of transformer-based models for sequential recommendation. We empirically evaluate RETR on seven intra-domain benchmarks and RETR yields state-of-the-art performance. On another five cross-domain benchmarks, RETR can capture more domain-invariant representations for sequential recommendation.

****

Towards Discovering Neural Architectures from Scratch

Simon Schrodi,Danny Stoll,Binxin Ru,Rhea Sanjay Sukthanker,Thomas Brox,Frank Hutter

The discovery of neural architectures from scratch is the long-standing goal of Neural Architecture Search (NAS). Searching over a wide spectrum of neural architectures can facilitate the discovery of previously unconsidered but well-performing architectures. In this work, we take a large step towards discovering neural architectures from scratch by expressing architectures algebraically. This algebraic view leads to a more general method for designing search spaces, which allows us to compactly represent search spaces that are 100s of orders of magnitud e larger than common spaces from the literature. Further, we propose a Bayesian Optimization strategy to efficiently search over such huge spaces, and demonstrate empirically that both our search space design and our search strategy can be superior to existing baselines. We open source our algebraic NAS approach and provide APIs for PyTorch and TensorFlow.

****

Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs

Yi-Lun Liao,Tess Smidt

Despite their widespread success in various domains, Transformer networks have yet to perform well across datasets in the domain of 3D atomistic graphs such as molecules even when 3D-related inductive biases like translational invariance and rotational equivariance are considered. In this paper, we demonstrate that Transformers can generalize well to 3D atomistic graphs and present Equiformer, a graph neural network leveraging the strength of Transformer architectures and incorporating SE(3)/E(3)-equivariant features based on irreducible representations (irreps). First, we propose a simple and effective architecture by only replacing original operations in Transformers with their equivariant counterparts and including tensor products. Using equivariant operations enables encoding equivariant information in channels of irreps features without complicating graph structures. With minimal modifications to Transformers, this architecture has already achieved strong empirical results. Second, we propose a novel attention mechanism called equivariant graph attention, which improves upon typical attention in Transformers through replacing dot product attention with multi-layer perceptron attention and including non-linear message passing. With these two innovations, Equiformer achieves competitive results to previous models on QM9, MD17 and OC20 datasets.

****

Automating Nearest Neighbor Search Configuration with Constrained Optimization

Philip Sun,Ruiqi Guo,Sanjiv Kumar

The approximate nearest neighbor (ANN) search problem is fundamental to efficiently serving many real-world machine learning applications. A number of techniques have been developed for ANN search that are efficient, accurate, and scalable. However, such techniques typically have a number of parameters that affect the speed-recall tradeoff, and exhibit poor performance when such parameters aren't properly set. Tuning these parameters has traditionally been a manual process, demanding in-depth knowledge of the underlying search algorithm. This is becoming an increasingly unrealistic demand as ANN search grows in popularity. To tackle this obstacle to ANN adoption, this work proposes a constrained optimization-based approach to tuning quantization-based ANN algorithms. Our technique takes just a desired search cost or recall as input, and then generates tunings that, empirically, are very close to the speed-recall Pareto frontier and give leading performance on standard benchmarks.

****

Truncated Diffusion Probabilistic Models and Diffusion-based Adversarial Auto-Encoders

Huangjie Zheng,Pengcheng He,Weizhu Chen,Mingyuan Zhou

Employing a forward diffusion chain to gradually map the data to a noise distribution, diffusion-based generative models learn how to generate the data by inferring a reverse diffusion chain. However, this approach is slow and costly because it needs many forward and reverse steps. We propose a faster and cheaper approach that adds noise not until the data become pure random noise, but until they reach a hidden noisy data distribution that we can confidently learn. Then, we use fewer reverse steps to generate data by starting from this hidden distribution that is made similar to the noisy data. We reveal that the proposed model can be cast as an adversarial auto-encoder empowered by both the diffusion process and a learnable implicit prior. Experimental results show even with a significantly smaller number of reverse diffusion steps, the proposed truncated diffusion probabilistic models can provide consistent improvements over the non-truncated ones in terms of performance in both unconditional and text-guided image generations.

****

NTK-SAP: Improving neural network pruning by aligning training dynamics

Yite Wang,Dawei Li,Ruoyu Sun

Pruning neural networks before training has received increasing interest due to its potential to reduce training time and memory. One popular method is to prune the connections based on a certain metric, but it is not entirely clear what metric is the best choice. Recent advances in neural tangent kernel (NTK) theory suggest that the training dynamics of large enough neural networks is closely related to the spectrum of the NTK. Motivated by this finding, we propose to prune the connections that have the least influence on the spectrum of the NTK. This method can help maintain the NTK spectrum, which may help align the training dynamics to that of its dense counterpart. However, one possible issue is that the fixed-weight-NTK corresponding to a given initial point can be very different from the NTK corresponding to later iterates during the training phase. We further propose to sample multiple realizations of random weights to estimate the NTK spectrum. Note that our approach is weight-agnostic, which is different from most existing methods that are weight-dependent. In addition, we use random inputs to compute the fixed-weight-NTK, making our method data-agnostic as well. We name our foresight pruning algorithm Neural Tangent Kernel Spectrum-Aware Pruning (NTK-SAP). Empirically, our method achieves better performance than all baselines on multiple datasets.

****

Towards Equivariant Graph Contrastive Learning via Cross-Graph Augmentation

Zhiyuan Liu,An Zhang,Yu Sun,Yicong Li,Yaorui Shi,Sihang Li,Xiang Wang,Xiangnan He,Tat-Seng Chua

Leading graph contrastive learning (GCL) frameworks conform to the invariance mechanism by encouraging insensitivity to different augmented views of the same graph. Despite the promising performance, invariance worsens representation when augmentations cause aggressive semantics shifts. For example, dropping the super-node can dramatically change a social network's topology. In this case, encouraging invariance to the original graph can bring together dissimilar patterns and hurt the task of instance discrimination. To resolve the problem, we get inspiration from equivariant self-supervised learning and propose Equivariant Graph Contrastive Learning (E-GCL) to encourage the sensitivity to global semantic shifts. Viewing each graph as a transformation to others, we ground the equivariance principle as a cross-graph augmentation -- graph interpolation -- to simulate global semantic shifts. Without using annotation, we supervise the representation of cross-graph augmented views by linearly combining the representations of their original samples. This simple but effective equivariance principle empowers E-GCL with the ability of cross-graph discrimination. It shows significant improvements over the state-of-the-art GCL models in unsupervised learning and transfer learning. Further experiments demonstrate E-GCL's generalization to various graph pre-training frameworks. Code is available at \url{https://anonymous.4open.science/r/E-GCL/}

****

Effective Self-supervised Pre-training on Low-compute Networks without Distillation

Fuwen Tan,Fatemeh Sadat Saleh,Brais Martinez

Despite the impressive progress of self-supervised learning (SSL), its applicability to low-compute networks has received limited attention. Reported performance has trailed behind standard supervised pre-training by a large margin, barring self-supervised learning from making an impact on models that are deployed on device. Most prior works attribute this poor performance to the capacity bottleneck of the low-compute networks and opt to bypass the problem through the use of knowledge distillation (KD). In this work, we revisit SSL for efficient neural networks, taking a closer at what are the detrimental factors causing the practical limitations, and whether they are intrinsic to the self-supervised low-compute setting. We find that, contrary to accepted knowledge, there is no intrinsic architectural bottleneck, we diagnose that the performance bottleneck is related to the model complexity vs regularization strength trade-off. In particular, we start by empirically observing that the use of local views can have a dramatic impact on the effectiveness of the SSL methods. This hints at view sampling being one of the performance bottlenecks for SSL on low-capacity networks. We hypothesize that the view sampling strategy for large neural networks, which requires matching views in very diverse spatial scales and contexts, is too demanding for low-capacity architectures. We systematize the design of the view sampling mechanism, leading to a new training methodology that consistently improves the performance across different SSL methods (e.g. MoCo-v2, SwAV or DINO), different low-size networks (convolution-based networks, e.g. MobileNetV2, ResNet18, ResNet34 and vision transformer, e.g. ViT-Ti), and different tasks (linear probe, object detection, instance segmentation and semi-supervised learning). Our best models establish new state-of-the-art for SSL methods on low-compute networks despite not using a KD loss term. Code is publicly available at github.com/saic-fi/SSLight.

****

Graph Neural Bandits

Yunzhe Qi,Yikun Ban,Jingrui He

Contextual bandits aim to choose the optimal arm with the highest reward out of a set of candidates based on their contextual information, and various bandit algorithms have been applied to personalized recommendation due to their ability of solving the exploitation-exploration dilemma. Motivated by online recommendation scenarios, in this paper, we propose a framework named Graph Neural Bandits (GNB) to leverage the collaborative nature among users empowered by graph neural networks (GNNs). Instead of estimating rigid user clusters, we model the "fine-grained'' collaborative effects through estimated user graphs in terms of exploitation and exploration individually. Then, to refine the recommendation strategy, we utilize separate GNN-based models on estimated user graphs for exploitation and adaptive exploration. Theoretical analysis and experimental results on multiple real data sets in comparison with state-of-the-art baselines are provided to demonstrate the effectiveness of our proposed framework.

****

CoRTX: Contrastive Framework for Real-time Explanation

Yu-Neng Chuang,Guanchu Wang,Fan Yang,Quan Zhou,Pushkar Tripathi,Xuanting Cai,Xia Hu

Recent advancements in explainable machine learning provide effective and faithful solutions for interpreting model behaviors. However, many explanation methods encounter efficiency issues, which largely limit their deployments in practical scenarios. Real-time explainer (RTX) frameworks have thus been proposed to accelerate the model explanation process by learning an one-feed-forward explainer. Existing RTX frameworks typically build the explainer under the supervised learning paradigm, which requires large amounts of explanation labels as the ground truth. Considering that accurate explanation labels are usually hard to obtain, due to constrained computational resources and limited human efforts, effective explainer training is still challenging in practice. In this work, we propose a COntrastive Real-Time eXplanation (CoRTX) framework to learn the explanation-oriented representation and relieve the intensive dependence of explainer training on explanation labels. Specifically, we design a synthetic strategy to select positive and negative instances for explanation representation learning. Theoretical analysis show that our selection strategy can benefit the contrastive learning process on explanation tasks. Experimental results on three real-world datasets further demonstrate the efficiency and efficacy of our proposed CoRTX framework.

****

MPCFORMER: FAST, PERFORMANT AND PRIVATE TRANSFORMER INFERENCE WITH MPC

Dacheng Li,Hongyi Wang,Rulin Shao,Han Guo,Eric Xing,Hao Zhang

Enabling private inference is crucial for many cloud inference services that are based on Transformer models. However, existing private inference solutions can increase the inference latency by more than 60$\times$ or significantly compromise the inference quality. In this paper, we design the framework MPCFORMER as a practical solution, using Secure Multi-Party Computation (MPC) and Knowledge Dis tillation (KD). Through extensive evaluations, we show that MPCFORMER significantly speeds up Transformer inference in MPC settings while achieving similar ML performance to the input model. On the IMDb dataset, it achieves similar performance to BERT_BASE, while being 5.3$\times$ faster. On the GLUE benchmark, it achieves 97% performance of BERT_BASE with a 2.2$\times$ speedup. MPCFORMER remains effective with different trained Transformer weights such as $\text{ROBERTA}_\text{BASE}$ and larger models including BERT_LARGE. Code is available at https://github.com/MccRee177/MPCFormer.

****

Discovering Distinctive ``Semantics'' in Super-Resolution Networks

Yihao Liu,Anran Liu,Jinjin Gu,Zhipeng Zhang,Wenhao Wu,Yu Qiao,Chao Dong

Image super-resolution (SR) is a representative low-level vision problem. Although deep SR networks have achieved extraordinary success, we are still unaware of their working mechanisms. Specifically, whether SR networks can learn semantic information, or just perform complex mapping function? What hinders SR networks from generalizing to real-world data? These questions not only raise our curiosity, but also influence SR network development. In this paper, we make the primary attempt to answer the above fundamental questions. After comprehensively analyzing the feature representations (via dimensionality reduction and visualization), we successfully discover the distinctive ``semantics'' in SR networks, i.e., deep degradation representations (DDR), which relate to image degradation instead of image content. We show that a well-trained deep SR network is naturally a good descriptor of degradation information. Our experiments also reveal two key factors (adversarial learning and global residual) that influence the extraction of such semantics. We further apply DDR in several interesting applications (such as distortion identification, blind SR and generalization evaluation) and achieve promising results, demonstrating the correctness and effectiveness of our findings.

****

Networks are Slacking Off: Understanding Generalization Problem in Image Deraining

Jinjin Gu,Xianzheng Ma,Xiangtao Kong,Yu Qiao,Chao Dong

Deep low-level networks are successful in laboratory benchmarks, but still suffer from severe generalization problems in real-world applications, especially for the deraining task. An ``acknowledgement'' of deep learning drives us to use the training data with higher complexity, expecting the network to learn richer knowledge to overcome generalization problems. Through extensive systematic experiments, we show that this approach fails to improve their generalization ability but instead makes the networks overfit to degradations even more. Our experiments establish that it is capable of training a deraining network with better generalization by reducing the training data complexity. Because the networks are slacking off during training, i.e. learn the less complex element in the image content and degradation to reduce the training loss. When the background image is less complex than the rain streak, the network will focus on the reconstruction of the background without overfitting the rain patterns, thus achieving a good generalization effect. Our research demonstrates excellent application potential and provides an indispensable perspective and research methodology for understanding the generalization problem of low-level vision.

****

Disparate Impact in Differential Privacy from Gradient Misalignment

Maria S. Esipova,Atiyeh Ashari Ghomi,Yaqiao Luo,Jesse C Cresswell

As machine learning becomes more widespread throughout society, aspects including data privacy and fairness must be carefully considered, and are crucial for deployment in highly regulated industries. Unfortunately, the application of privacy enhancing technologies can worsen unfair tendencies in models. In particular, one of the most widely used techniques for private model training, differentially private stochastic gradient descent (DPSGD), frequently intensifies disparate impact on groups within data. In this work we study the fine-grained causes of unfairness in DPSGD and identify gradient misalignment due to inequitable gradie nt clipping as the most significant source. This observation leads us to a new method for reducing unfairness by preventing gradient misalignment in DPSGD.

****

IDP: Iterative Differentiable Pruning based on Attention for Deep Neural Networks

Minsik Cho,Saurabh Adya,Devang Naik

Deep Neural network (DNN) pruning is an effective method to reduce the size of a model, improve the inference latency, and minimize the power consumption on DNN accelerators, at the risk of decreasing model accuracy. In this paper, we propose a novel differentiable pruning scheme, Iterative Differentiable Pruning or IDP which offers state-of-the-art qualities in model size, accuracy, and training cost. IDP creates attention-based pruning masks for a given sparsity target to achieve the state-of-the-art trade-offs between model accuracy and inference compute with negligible training overhead. We evaluated IDP on various computer vision and natural language processing tasks, and found that IDP delivers the state-of-the-art results. For MobileNet-v1 (which is a challenging DNN for pruning), IDP can achieve 68.2% top-1 ImageNet1k accuracy with 86.6% sparsity which is 2.3% higher accuracy than the latest state-of-the-art pruning algorithms. For ResNet18, IDP offers 69.5% top-1 ImageNet1k accuracy with 85.5% sparsity at the same training budget and 0.8% better top-1 accuracy than the state-of-the-art method. Also, IDP demonstrates over 83.1% accuracy on Multi-Genre Natural Language Inference with 90% sparsity for BERT, while the next best from the existing techniques shows 81.5% accuracy.

****

FADE: Enabling Large-Scale Federated Adversarial Training on Resource-Constrained Edge Devices

Minxue Tang,Jianyi Zhang,Mingyuan Ma,Louis DiValentin,Aolin Ding,Amin Hassanzadeh,Hai Li,Yiran Chen

Federated adversarial training can effectively complement adversarial robustness into the privacy-preserving federated learning systems. However, the high demand for memory capacity and computing power makes large-scale federated adversarial training infeasible on resource-constrained edge devices. Few previous studies in federated adversarial training have tried to tackle both memory and computational constraints at the same time. In this paper, we propose a new framework named Federated Adversarial Decoupled Learning (FADE) to enable AT on resource-constrained edge devices. FADE decouples the entire model into small modules to fit into the resource budget of each edge device respectively, and each device only needs to perform AT on a single module in each communication round. We also propose an auxiliary weight decay to alleviate objective inconsistency and achieve better accuracy-robustness balance in FADE. FADE offers a theoretical guarantee for convergence and adversarial robustness, and our experimental results show that FADE can significantly reduce the consumption of memory and computing power while maintaining accuracy and robustness.

****

Temporal Relevance Analysis for Video Action Models

Quanfu Fan,Donghyun Kim,Chun-Fu Chen,Stan Sclaroff,Kate Saenko,Sarah Adel Bargal

In this paper, we provide a deep analysis of temporal modeling for action recognition, an important but underexplored problem in
the literature. We first propose a new approach to quantify the temporal relationships between frames captured by CNN-based action models based on layer-wise relevance propagation. We then conduct comprehensive experiments and in-depth analysis to provide a better understanding of how temporal modeling is affected by various factors such as dataset, network architecture, and input frames. With this, we further study some important questions for action recognition that lead to interesting findings. Our analysis shows that there is no strong correlation between temporal relevance and model performance; and action models tend to capture local temporal information, but less long-range dependencies.

****

DeepPipe: Deep, Modular and Extendable Representations of Machine Learning Pipelines

Sebastian Pineda Arango,Josif Grabocka

Finding accurate Machine Learning pipelines is essential in achieving state-of-the-art AI predictive performance. Unfortunately, most existing Pipeline Optimization techniques rely on flavors of Bayesian Optimization that do not explore the deep interaction between pipeline stages/components (e.g. between hyperparameters of the deployed preprocessing algorithm and the hyperparameters of a classifier). In this paper, we are the first to capture the deep interaction between components of a Machine Learning pipeline. We propose embedding pipelines in a deep latent representation through a novel per-component encoder mechanism. Such pipeline embeddings are used with deep kernel Gaussian Process surrogates inside a Bayesian Optimization setup. Through extensive experiments on large-scale meta-datasets, we demonstrate that learning pipeline embeddings with Deep Neural Networks significantly advances the state-of-the-art in Pipeline Optimization.

****

OTOv2: Automatic, Generic, User-Friendly

Tianyi Chen,Luming Liang,Tianyu DING,Zhihui Zhu,Ilya Zharkov

The existing model compression methods via structured pruning typically require complicated multi-stage procedures. Each individual stage necessitates numerous engineering efforts and domain-knowledge from the end-users which prevent their wider applications onto broader scenarios. We propose the second generation of Only-Train-Once (OTOv2), which first automatically trains and compresses a general DNN only once from scratch to produce a more compact model with competitive performance without fine-tuning. OTOv2 is automatic and pluggable into various deep learning applications, and requires almost minimal engineering efforts from the users. Methodologically, OTOv2 proposes two major improvements: (i) Autonomy: automatically exploits the dependency of general DNNs, partitions the trainable variables into Zero-Invariant Groups (ZIGs), and constructs the compressed model; and (ii) Dual Half-Space Projected Gradient (DHSPG): a novel optimizer to more reliably solve structured-sparsity problems. Numerically, we demonstrate the generality and autonomy of OTOv2 on a variety of model architectures such as VGG, ResNet, CARN, ConvNeXt, DenseNet and StackedUnets, the majority of which cannot be handled by other methods without extensive handcrafting efforts. Together with benchmark datasets including CIFAR10/100, DIV2K, Fashion-MNIST, SVNH and ImageNet, its effectiveness is validated by performing competitively or even better than the state-of-the-arts. The source code is available at https://github.com/tianyic/only_train_once.

****

TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second

Noah Hollmann,Samuel Müller,Katharina Eggensperger,Frank Hutter

We present TabPFN, a trained Transformer that can do supervised classification for small tabular datasets in less than a second, needs no hyperparameter tuning and is competitive with state-of-the-art classification methods.

TabPFN is fully entailed in the weights of our network, which accepts training and test samples as a set-valued input and yields predictions for the entire test set in a single forward pass.

TabPFN is a Prior-Data Fitted Network (PFN) and is trained offline once, to approximate Bayesian inference on synthetic datasets drawn from our prior.

This prior incorporates ideas from causal reasoning: It entails a large space of structural causal models with a preference for simple structures.

On the 18 datasets in the OpenML-CC18 suite that contain up to 1000 training data points, up to 100 purely numerical features without missing values, and up to 10 classes, we show that our method clearly outperforms boosted trees and performs on par with complex state-of-the-art AutoML systems with up to $230\times$ speedup.

This increases to a $5\,700\times$ speedup when using a GPU. We also validate these results on an additional 67 small numerical datasets from OpenML.

We provide all our code, the trained TabPFN, an interactive browser demo and a Colab notebook at https://github.com/automl/TabPFN.

****

On the Importance of Architectures and Hyperparameters for Fairness in Face Recognition
Rhea Sanjay Sukthanker,Samuel Dooley,John P Dickerson,Colin White,Frank Hutter,Micah Goldblum
Face recognition systems are deployed across the world by government agencies and contractors for sensitive and impactful tasks, such as surveillance and database matching. Despite their widespread use, these systems are known to exhibit bias across a range of sociodemographic dimensions, such as gender and race. Nonetheless, an array of works proposing pre-processing, training, and post-processing methods have failed to close these gaps. Here, we take a very different approach to this problem, identifying that both architectures and hyperparameters of neural networks are instrumental in reducing bias. We first run a large-scale analysis of the impact of architectures and training hyperparameters on several common fairness metrics and show that the implicit convention of choosing high-accuracy architectures may be suboptimal for fairness. Motivated by our findings, we run the first neural architecture search for fairness, jointly with a search for hyperparameters. We output a suite of models which Pareto-dominate all other competitive architectures in terms of accuracy and fairness. Furthermore, we show that these models transfer well to other face recognition datasets with similar and distinct protected attributes. We release our code and raw result files so that researchers and practitioners can replace our fairness metrics with a bias measure of their choice.
****
Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data
Yaoqing Yang,Ryan Theisen,Liam Hodgkinson,Joseph E. Gonzalez,Kannan Ramchandran,charles h martin,Michael W. Mahoney
The search for effective and robust metrics has been the focus of recent theoretical and empirical work on generalization of deep neural networks (NNs). In this paper, we discuss the performance of natural language processing (NLP) models, and we evaluate various existing and novel generalization metrics. Compared to prior studies, we (i) focus on NLP instead of computer vision (CV), (ii) focus on generalization metrics that predict test error instead of the generalization gap, (iii) focus on generalization metrics that do not need the access to data, and (iv) focus on the heavy-tail (HT) phenomenon that has received comparatively less attention in the study of deep neural networks. We extend recent HT-based work which focuses on power law (PL) distributions, and we study exponential (EXP) and exponentially truncated power law (E-TPL) fitting to the empirical spectral densities (ESDs) of weight matrices. Our empirical studies are carried on (i) hundreds of Transformers trained in different settings, in which we systematically vary the amount of data, the model size and the optimization hyperparameters, (ii) a total of 51 pretrained Transformers from eight families of Huggingface NLP models, including BERT, GPT2, ALBERT, etc., and (iii) a total of 28 existing and novel generalization metrics. From our detailed empirical analyses, we show that shape metrics, or the metrics obtained from fitting the shape of the ESDs, perform uniformly better at predicting generalization performance than scale metrics commonly studied in the literature, as measured by the average rank correlations with the generalization performance for all of our experiments. We also show that among the three HT distributions considered in our paper, the E-TPL fitting of ESDs performs the most robustly when the models are trained in experimental settings, while the PL fitting achieves the best performance on well-trained Huggingface models, and that both E-TPL and PL metrics (which are both shape metrics) outperform scale metrics.
****
Human Motion Diffusion Model
Guy Tevet,Sigal Raab,Brian Gordon,Yoni Shafir,Daniel Cohen-or,Amit Haim Bermano
Natural and expressive human motion generation is the holy grail of computer animation.
It is a challenging task, due to the diversity of possible motion, human perceptual sensitivity to it, and the difficulty of accurately describing it. Therefore

, current generative solutions are either low-quality or limited in expressiveness.
Diffusion models are promising candidates for the human motion domain since they have already shown remarkable generative capabilities in other domains, and their many-to-many nature.
In this paper, we introduce Motion Diffusion Model (MDM), a carefully adapted classifier-free diffusion-based generative model for human motion data. MDM is transformer-based, combining insights from motion generation literature.
A notable design-choice is that it predicts the sample itself rather than the noise in each step to facilitate the use of established geometric losses on the locations and velocities of the motion, such as the foot contact loss. As we demonstrate, MDM is a generic approach, enabling different modes of conditioning, and different generation tasks. We show that our model is trained with lightweight resources and yet achieves state-of-the-art results on leading benchmarks for text-to-motion, action-to-motion, and unconditioned motion generation.

****

Structure-based Drug Design with Equivariant Diffusion Models
Arne Schneuing,Yuanqi Du,Charles Harris,Arian Rokkum Jamasb,Ilia Igashov,weitao Du,Tom Leon Blundell,Pietro Lio,Carla P Gomes,Max Welling,Michael M. Bronstein,Bruno Correia
Structure-based drug design (SBDD) aims to design small-molecule ligands that bind with high affinity and specificity to pre-determined protein targets. Traditional SBDD pipelines start with large-scale docking of compound libraries from public databases, thus limiting the exploration of chemical space to existent previously studied regions.
Recent machine learning methods approached this problem using an atom-by-atom generation approach, which is computationally expensive.
In this paper, we formulate SBDD as a 3D-conditional generation problem and present DiffSBDD, an E(3)-equivariant 3D-conditional diffusion model that generates novel ligands conditioned on protein pockets.
Furthermore, we curate a new dataset of experimentally determined binding complex data from Binding MOAD to provide realistic binding scenario rather than the synthetic CrossDocked dataset. Comprehensive in silico experiments demonstrate the efficiency of DiffSBDD in generating novel and diverse drug-like ligands that engage protein pockets with high binding energies as predicted by in silico docking.

****

Deep reinforced active learning for multi-class image classification
Emma Slade,Kim Branson
High accuracy medical image classification can be limited by the costs of acquiring more data as well as the time and expertise needed to label existing images. In this paper, we apply active learning to medical image classification, a method which aims to maximise model performance on a minimal subset from a larger pool of data. We present a new active learning framework, based on deep reinforcement learning, to learn an active learning query strategy to label images based on predictions from a convolutional neural network. Our framework modifies the deep-Q network formulation, allowing us to pick data based additionally on geometric arguments in the latent space of the classifier, allowing for high accuracy multi-class classification in a batch-based active learning setting, enabling the agent to label datapoints that are both diverse and about which it is most uncertain. We apply our framework to two medical imaging datasets and compare with standard query strategies as well as the most recent reinforcement learning based active learning approach for image classification.

****

A UNIFIED VIEW OF FINDING AND TRANSFORMING WINNING LOTTERY TICKETS
Kun Wang,Yuxuan Liang,Pengkun Wang,Pengfei Gu,Zhengyang Zhou,Chao Huang,Yang Wang
While over-parameterized deep neural networks obtain prominent results on various machine learning tasks, their superfluous parameters usually make model training and inference notoriously inefficient. Lottery Ticket Hypothesis (LTH) addres

ses this issue from a novel perspective: it articulates that there always exist sparse and admirable subnetworks in a randomly initialized dense network, which can be realized by an iterative pruning strategy. Dual Lottery Ticket Hypothesis (DLTH) further investigates sparse network training from a complementary view. Concretely, it introduces a gradually increased regularization term to transform a dense network to an ultra-light subnetwork without sacrificing learning capacity. After revisiting the success of LTH and DLTH, we unify these two research lines by coupling the stability of iterative pruning and the excellent performance of increased regularization, resulting in two new algorithms (UniLTH and UniDLTH) for finding and transforming winning tickets, respectively. Unlike either LTH without regularization or DLTH which applies regularization across the training, our methods first train the network without any regularization force until the model reaches a certain point (i.e., the validation loss does not decrease for several epochs), and then employ increased regularization for information extrusion and iteratively perform magnitude pruning till the end. We theoretically prove that the early stopping mechanism acts analogously as regularization and can help the optimization trajectory stop at a particularly better point in space than regularization. This not only prevent the parameters from being excessively skewed to the training distribution (over-fitting), but also better stimulate the network potential to obtain more powerful subnetworks. Extensive experiments are conducted to show the superiority of our methods in terms of accuracy and sparsity.

****

Filter-Recovery Network for Multi-Speaker Audio-Visual Speech Separation

Haoyue Cheng,Zhaoyang Liu,Wayne Wu,Limin Wang

In this paper, we systematically study the audio-visual speech separation task in a multi-speaker scenario. Given the facial information of each speaker, the goal of this task is to separate the corresponding speech from the mixed speech. The existing works are designed for speech separation in a controlled setting with a fixed number of speakers (mostly 2 or 3 speakers), which seems to be impractical for real applications. As a result, we try to utilize a single model to separate the voices with a variable number of speakers. Based on the observation, there are two prominent issues for multi-speaker separation: 1) There are some noisy voice pieces belonging to other speakers in the separation results; 2) Part of the target speech is missing after separation. Accordingly, we propose \textbf{BFRNet}, including a {\bf B}asic audio-visual speech separator and a Filter-Recovery Network (\textbf{FRNet}). FRNet can refine the coarse audio separated by basic audio-visual speech separator. To have fair comparisons, we build a comprehensive benchmark for multi-speaker audio-visual speech separation to verify the performance of various methods. Experimental results show that our method is able to achieve the state-of-the-art performance. Furthermore, we also find that FRNet can boost the performance of other off-the-shelf speech separators, which exhibits its ability of generalization.

****

Probing into the Fine-grained Manifestation in Multi-modal Image Synthesis

Qianyu Feng,Peike Li,Yulei Sui,Hongyu Zhang

The ever-growing development of multi-modal image synthesis brings unprecedented realism to generation tasks. In practice, it is straightforward to judge the visual quality and reality of an image. However, it is labor-consuming to verify the correctness of semantic consistency in the auto-generation, which requires a comprehensive understanding and mapping of different modalities. The results of existing models are sorted and displayed largely relying on the global visual-text similarity. However, this coarse-grained approach does not capture the fine-grained semantic alignment between image regions and text spans. To address this issue, we first present a new method to evaluate the cross-modal consistency by inspecting the decomposed semantic concepts. We then introduce a new metric, called MIS-Score, which is designed to measure the fine-grained semantic alignment between a prompt and its generation quantitatively. Moreover, we have also developed an automated robustness testing technique with referential transforms to test and measure the robustness of multi-modal synthesis models. We have conducte d comprehensive experiments to evaluate the performance of recent popular models for text-to-image generation. Our study demonstrates that the proposed metric MIS-Score represents better evaluation criteria than existing coarse-grained ones (e.g., CLIP) to understand the semantic consistency of the synthesized results. Our robustness testing method also proves the existence of biases embedded in the models, hence uncovering their limitations in real applications.

****

Can discrete information extraction prompts generalize across language models?

Nathanaël Carraz Rakotonirina,Roberto Dessi,Fabio Petroni,Sebastian Riedel,Marco Baroni

We study whether automatically-induced prompts that effectively extract information from a language model can also be used, out-of-the-box, to probe other language models for the same information. After confirming that discrete prompts induced with the AutoPrompt algorithm outperform manual and semi-manual prompts on the slot-filling task, we demonstrate a drop in performance for AutoPrompt prompts learned on a model and tested on another. We introduce a way to induce prompts by mixing language models at training time that results in prompts that generalize well across models. We conduct an extensive analysis of the induced prompts, finding that the more general prompts include a larger proportion of existing English words and have a less order-dependent and more uniform distribution of information across their component tokens. Our work provides preliminary evidence that it's possible to generate discrete prompts that can be induced once and used with a number of different models, and gives insights on the properties characterizing such prompts.

****

Deep Power Laws for Hyperparameter Optimization

Arlind Kadra,Maciej Janowski,Martin Wistuba,Josif Grabocka

Hyperparameter optimization is an important subfield of machine learning that focuses on tuning the hyperparameters of a chosen algorithm to achieve peak performance. Recently, there has been a stream of methods that tackle the issue of hyperparameter optimization, however, most of the methods do not exploit the scaling law property of learning curves. In this work, we propose Deep Power Law (DPL), a neural network model conditioned to yield predictions that follow a power-law scaling pattern. Our model dynamically decides which configurations to pause and train incrementally by making use of multi-fidelity estimation. We compare our method against 7 state-of-the-art competitors on 3 benchmarks related to tabular, image, and NLP datasets covering 59 diverse search spaces. Our method achieves the best results across all benchmarks by obtaining the best any-time results compared to all competitors.

****

A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta.

Maksim Velikanov,Denis Kuznedelev,Dmitry Yarotsky

Mini-batch SGD with momentum is a fundamental algorithm for learning large predictive models. In this paper we develop a new analytic framework to analyze noise-averaged properties of mini-batch SGD for linear models at constant learning rates, momenta and sizes of batches. Our key idea is to consider the dynamics of the second moments of model parameters for a special family of "Spectrally Expressible" approximations. This allows to obtain an explicit expression for the generating function of the sequence of loss values. By analyzing this generating function, we find, in particular, that 1) the SGD dynamics exhibits several convergent and divergent regimes depending on the spectral distributions of the problem; 2) the convergent regimes admit explicit stability conditions, and explicit loss asymptotics in the case of power-law spectral distributions; 3) the optimal convergence rate can be achieved at negative momenta. We verify our theoretical predictions by extensive experiments with MNIST and synthetic problems, and find a good quantitative agreement.

****

Curiosity-Driven Unsupervised Data Collection for Offline Reinforcement Learning

Chenyu Sun,Hangwei Qian,Chunyan Miao

In offline reinforcement learning (RL), while the majority of efforts are focusing on engineering sophisticated learning algorithms given a fixed dataset, very few works have been carried out to improve the dataset quality itself. More importantly, it is even challenging to collect a task-agnostic dataset such that the offline RL agent can learn multiple skills from it. In this paper, we propose a Curiosity-driven Unsupervised Data Collection (CUDC) method to improve the data collection process. Specifically, we quantify the agent's internal belief to estimate the probability of the k-step future states being reachable from the current states. Different from existing approaches that implicitly assume limited feature space with fixed environment steps, CUDC is capable of adapting the number of environment steps to explore. Thus, the feature representation can be substantially diversified with the dynamics information. With this adaptive reachability mechanism in place, the agent can navigate itself to collect higher-quality data with curiosity. Empirically, CUDC surpasses existing unsupervised methods in sample efficiency and learning performance in various downstream offline RL tasks of the DeepMind control suite.

****

Big Learning: A Universal Machine Learning Paradigm?

Yulai Cong,Miaoyun Zhao

Recent breakthroughs based on big/foundation models reveal a vague avenue for AI, that is, \emph{big data, big/foundation models, big learning, $\cdots$}. Following that avenue, here we elaborate on our newly introduced big learning. Specifically, big learning exhaustively exploits the information/tasks inherent in its large-scale \emph{complete/incomplete} training data, by learning to simultaneously model many/all joint/conditional/marginal data distributions (thus named big learning) with one universal foundation model. We reveal that big learning is what existing foundation models are implicitly doing; accordingly, our big learning provides high-level guidance for flexible design and improvements of foundation models. Besides, big learning (i) is equipped with great flexibilities for complete/incomplete training data and for customizing trustworthy data tasks; (ii) potentially delivers all joint/conditional/marginal data capabilities after training; (iii) significantly reduces the training-test gap with improved model generalization; and (iv) potentially unifies conventional machine learning paradigms and enables their flexible cooperations, manifested as a universal learning paradigm. Preliminary experiments verified the effectiveness of the presented big learning.

****

Bridging the Gap between ANNs and SNNs by Calibrating Offset Spikes

Zecheng Hao,Jianhao Ding,Tong Bu,Tiejun Huang,Zhaofei Yu

Spiking Neural Networks (SNNs) have attracted great attention due to their distinctive characteristics of low power consumption and temporal information processing. ANN-SNN conversion, as the most commonly used training method for applying SNNs, can ensure that converted SNNs achieve comparable performance to ANNs on large-scale datasets. However, the performance degrades severely under low quantities of time-steps, which hampers the practical applications of SNNs to neuromorphic chips.

In this paper, instead of evaluating different conversion errors and then eliminating these errors, we define an offset spike to measure the degree of deviation between actual and desired SNN firing rates. We perform a detailed analysis of offset spike and note that the firing of one additional (or one less) spike is the main cause of conversion errors. Based on this, we propose an optimization strategy based on shifting the initial membrane potential and we theoretically prove the corresponding optimal shifting distance for calibrating the spike. In addition, we also note that our method has a unique iterative property that enables further reduction of conversion errors. The experimental results show that our proposed method achieves state-of-the-art performance on CIFAR-10, CIFAR-100, and ImageNet datasets. For example, we reach a top-1 accuracy of 67.12% on ImageNet when using 6 time-steps. To the best of our knowledge, this is the first time an ANN-SNN conversion has been shown to simultaneously achieve high accuracy and ultralow latency on complex datasets. Code is available at https://github.com/hzc1208/ANN2SNN_COS.

****

MIA: A Framework for Certified Robustness of Time-Series Classification and Forecasting Against Temporally-Localized Perturbations

Ruoxin Chen,Ruizhe Zhong,Jiawei Sun,Jie LI,Chentao Wu,Junchi Yan

Recent literature demonstrates that times-series forecasting/classification are sensitive to input perturbations. However, the defenses for time-series models are relatively under-explored. In this paper, we propose \textbf{M}asking \textbf{I}mputing \textbf{A}ggregation (MIA), a plug-and-play framework to provide an arbitrary deterministic time-series model with certified robustness against temporally-localized perturbations (also known as ℓ_0-norm localized perturbations), which is to our knowledge the first ℓ_0-norm defense for time-series models. Our main insight is to let an occluding mask move across the input series, guaranteeing that, for an arbitrary localized perturbation there must exist at least a mask that completely remove out the perturbation, so that our prediction on this masked series is uninfluenced. Remarkably, MIA is high-availability as it still works even if we only have query access to the pretrained model. Furthermore, as there is no dedicated defense against ℓ_0-norm perturbations for time-series models, we specifically adapt two matrix-based defenses to time-series models for comparison. Extensive experiments show that MIA yields stronger robustness as well as practicality.

****

Offline RL with No OOD Actions: In-Sample Learning via Implicit Value Regularization

Haoran Xu,Li Jiang,Jianxiong Li,Zhuoran Yang,Zhaoran Wang,Victor Wai Kin Chan,Xianyuan Zhan

Most offline reinforcement learning (RL) methods suffer from the trade-off between improving the policy to surpass the behavior policy and constraining the policy to limit the deviation from the behavior policy as computing Q-values using out-of-distribution (OOD) actions will suffer from errors due to distributional shift. The recent proposed \textit{In-sample Learning} paradigm (i.e., IQL), which improves the policy by quantile regression using only data samples, shows great promise because it learns an optimal policy without querying the value function of any unseen actions. However, it remains unclear how this type of method handles the distributional shift in learning the value function. In this work, we make a key finding that the in-sample learning paradigm arises under the \textit{Implicit Value Regularization} (IVR) framework. This gives a deeper understanding of why the in-sample learning paradigm works, i.e., it applies implicit value regularization to the policy. Based on the IVR framework, we further propose two practical algorithms, Sparse Q-learning (SQL) and Exponential Q-learning (EQL), which adopt the same value regularization used in existing works, but in a complete in-sample manner. Compared with IQL, we find that our algorithms introduce sparsity in learning the value function, making them more robust in noisy data regimes. We also verify the effectiveness of SQL and EQL on D4RL benchmark datasets and show the benefits of in-sample learning by comparing them with CQL in small data regimes. Code is available at \url{https://github.com/ryanxhr/SQL}.

****

Eliminating Catastrophic Overfitting Via Abnormal Adversarial Examples Regularization

Runqi Lin,Chaojian Yu,Tongliang Liu

Single-step adversarial training (SSAT) is shown to be able to defend against iterative-step adversarial attacks to achieve both efficiency and robustness. However, SSAT suffers from catastrophic overfitting (CO) with strong adversaries, showing that the classifier decision boundaries are highly distorted and robust accuracy against iterative-step adversarial attacks suddenly drops from peak to nearly 0% in a few epochs. In this work, we find that some adversarial examples generated on the network trained by SSAT exhibit anomalous behaviour, that is, although the training data is generated by the inner maximization process, the loss of some adversarial examples decreases instead, which we called abnormal adversarial examples. Furthermore, network optimization on these abnormal adversarial examples will further accelerate the model decision boundaries distortion, and correspondingly, the number of abnormal adversarial examples will sharply increase with CO. These observations motivate us to prevent CO by hindering the generation of abnormal adversarial examples. Specifically, we design a novel method, Abnormal Adversarial Examples Regularization (AAER), which explicitly regularizes the number and logits variation of abnormal adversarial examples to hinder the model from generating abnormal adversarial examples. Extensive experiments demonstrate that our method can prevent CO and further boost adversarial robustness with strong adversaries.

****

TIB: Detecting Unknown Objects via Two-Stream Information Bottleneck

Aming WU,Cheng Deng

Detecting diverse objects, including ones never-seen-before during model training, is critical for the safe application of object detectors. To this end, a task of unsupervised out-of-distribution object detection (OOD-OD) is proposed to detect unknown objects without the reliance on an auxiliary dataset. For this task, it is important to reduce the impact of lacking unknown data for supervision and leverage in-distribution (ID) data to improve the model's discrimination ability. In this paper, we propose a method of Two-Stream Information Bottleneck (TIB), which consists of a standard Information Bottleneck and a dedicated Reverse Information Bottleneck (RIB). Specifically, after extracting the features of an ID image, we first define a standard IB network to disentangle instance representations that are beneficial for localizing and recognizing objects. Meanwhile, we present RIB to obtain simulative OOD features to alleviate the impact of lacking unknown data. Different from standard IB aiming to extract task-relevant compact representations, RIB is to obtain task-irrelevant representations by reversing the optimization objective of the standard IB. Next, to further enhance the discrimination ability, a mixture of information bottlenecks is designed to sufficiently capture object-related information. In the experiments, our method is evaluated on OOD-OD and incremental object detection. The significant performance gains over baselines show the superiorities of our method.

****

Win: Weight-Decay-Integrated Nesterov Acceleration for Adaptive Gradient Algorithms

Pan Zhou,Xingyu Xie,Shuicheng YAN

Training deep networks on large-scale datasets is computationally challenging. In this work, we explore the problem of ``\textit{how to accelerate adaptive gradient algorithms in a general manner}", and aim to provide practical efficiency-boosting insights.   To this end, we   propose an effective and general   {Weight-decay-Integrated Nesterov acceleration} (Win) to accelerate adaptive algorithms. Taking AdamW and Adam as examples, we minimize a dynamical   loss per iteration which combines the vanilla training loss and a dynamic regularizer inspired by proximal point method (PPM) to improve the convexity of the problem. To introduce Nesterov-alike-acceleration into AdamW and Adam, we respectively use the first- and second-order Taylor approximations of vanilla loss to update the variable twice. In this way, we arrive at our Win acceleration for AdamW and Adam that uses a conservative step and a reckless step to update twice and then linearly combines these two updates for acceleration. Next, we extend Win acceleration to LAMB and SGD. Our transparent acceleration derivation could provide insights for other accelerated methods and their integration into adaptive algorithms. Besides, we prove the convergence of Win-accelerated adaptive algorithms and justify their convergence superiority over their non-accelerated counterparts by taking AdamW and Adam as examples. Experimental results testify to the faster convergence speed and superior performance of our Win-accelerated AdamW, Adam, LAMB and SGD over their non-accelerated counterparts on vision classification tasks and language modeling tasks with both CNN and Transforme r backbones. We hope Win shall be a default acceleration option for popular optimizers in deep learning community to improve the training efficiency. Code will be released at \url{https://github.com/sail-sg/win}.

****

Towards Understanding Convergence and Generalization of AdamW

Pan Zhou,Xingyu Xie,Shuicheng YAN

AdamW modifies vanilla Adam by decaying network weights per training iteration, and shows remarkable generalization superiority over Adam and its ℓ_2-regularized variant. In context of adaptive gradient algorithms (\eg~Adam), the decoupled weight decay in AdamW differs from the widely used ℓ_2-regularizer, since the former does not affect optimization steps, while the latter changes the first- and second-order gradient moments and thus the optimization steps. Despite its great success on both vision transformers and CNNs, for AdamW, its convergence behavior and its generalization improvement over (ℓ_2-regularized) Adam remain absent yet. To solve this issue, we prove the convergence of AdamW and justify its generalization advantages over Adam and its ℓ_2-regularized version. Specifically, AdamW can provably converge but minimizes a dynamically regularized loss that combines a vanilla loss and a dynamical regularization induced by the decoupled weight decay, thus leading to its different behaviors compared with Adam and its ℓ_2-regularized version. Moreover, on both general nonconvex problems and P\L-conditioned problems, we establish the stochastic gradient complexity of AdamW to find a stationary point. Such complexity is also applicable to Adam and its ℓ_{2}-regularized variant, and indeed improves their previously known complexity, especially for modern over-parametrized networks. Besides, we theoretically show that AdamW often enjoys smaller generalization error bound than both Adam and its ℓ_2-regularized variant from the Bayesian posterior aspect. This result, for the first time, explicitly reveals the benefits of the unique decoupled weight decay in AdamW. We hope the theoretical results in this work could motivate researchers to propose novel optimizers with faster convergence and better generalization. Experimental results testify our theoretical implications.

****

GeoVeX: Geospatial Vectors with Hexagonal Convolutional Autoencoders

Daniele Donghi,Anne Morvan

We introduce a new geospatial representation model called GeoVeX to learn global vectors for all geographical locations on Earth land cover (200+ million embeddings). GeoVeX is built on a novel model architecture named Hexagonal Convolutional Autoencoders (HCAE) combined with a Zero-Inflated Poisson (ZIP) reconstruction layer, applied to a grid of Uber's H3 hexagons, each one described by the histogram of OpenStreetMap (OSM) geographical tags occurrences. GeoVeX is novel on three aspects: 1) it produces the first geospatial vectors trained on worldwide open data, enabling wide adoption on every downstream tasks which may benefit from enriched geographical information, requiring only location coordinates; 2) it represents the first use of hexagonal convolutions within autoencoder architectures, to learn latent representations of an hexagonal grid; and 3) it introduces a spatial-contextual Poisson reconstruction loss function for autoencoder architectures suitable for training on sparse geographical count data. Experiments demonstrate that GeoVeX embeddings can improve upon state-of-the-art geospatial location representations on two different downstream tasks: price prediction in the travel industry and hyperlocal interpolation of climate data from weather stations.

****

Split and Merge Proxy: pre-training protein-protein contact prediction by mining rich information from monomer data

Hao Du,Yuchen Ren,Yan Lu,He Huang,Yating Liu,Zhendong Mao,Xinqi Gong,Wanli Ouyang

Protein-protein contact prediction is a key intelligent biology computation technology for complex multimer protein function analysis but still sufferers from low accuracy. An important problem is that the number of training data cannot meet the requirements of deep-learning-based methods due to the expensive cost of c apturing structure information of multimer data. In this paper, we solve this data volume bottleneck in a cheap way, borrowing rich information from monomer data. To utilize monomer (single chain) data in this multimer (multiple chains) problem, we propose a simple but effective pre-training method called Split and Merger Proxy (SMP), which utilizes monomer data to construct a proxy task for model pre-training. This proxy task cuts monomer data into two sub-parts, called pseudo multimer, and pre-trains the model to merge them back together by predicting their pseudo contacts. The pre-trained model is then used to initialize for our target – protein-protein contact prediction. Because of the consistency between this proxy task and the final target, the whole method brings a stronger pre-trained model for subsequent fine-tuning, leading to significant performance gains. Extensive experiments validate the effectiveness of our method and show the model performs better than the state of the art by 11.40% and 2.97% on the P@ L/10 metric for bounded benchmarks DIPS-Plus and CASP-CAPRI, respectively. Further, the model also achieves almost 1.5 times performance superiority to the state of the art on the harder unbounded benchmark DB5. The code, model, and pre-training data will be released after this paper is accepted.

****

ESD: Expected Squared Difference as a Tuning-Free Trainable Calibration Measure

Hee Suk Yoon,Joshua Tian Jin Tee,Eunseop Yoon,Sunjae Yoon,Gwangsu Kim,Yingzhen Li,Chang D. Yoo

Studies have shown that modern neural networks tend to be poorly calibrated due to over-confident predictions. Traditionally, post-processing methods have been used to calibrate the model after training. In recent years, various trainable calibration measures have been proposed to incorporate them directly into the training process. However, these methods all incorporate internal hyperparameters, and the performance of these calibration objectives relies on tuning these hyperparameters, incurring more computational costs as the size of neural networks and datasets become larger. As such, we present Expected Squared Difference (ESD), a tuning-free (i.e., hyperparameter-free) trainable calibration objective loss, where we view the calibration error from the perspective of the squared difference between the two expectations. With extensive experiments on several architectures (CNNs, Transformers) and datasets, we demonstrate that (1) incorporating ESD into the training improves model calibration in various batch size settings without the need for internal hyperparameter tuning, (2) ESD yields the best-calibrated results compared with previous approaches, and (3) ESD drastically improves the computational costs required for calibration during training due to the absence of internal hyperparameter. The code is publicly accessible at https://github.com/hee-suk-yoon/ESD.

****

Interactive Portrait Harmonization

Jeya Maria Jose Valanarasu,HE Zhang,Jianming Zhang,Yilin Wang,Zhe Lin,Jose Echevarria,Yinglan Ma,Zijun Wei,Kalyan Sunkavalli,Vishal Patel

Current image harmonization methods consider the entire background as the guidance for harmonization. However, this may limit the capability for user to choose any specific object/person in the background to guide the harmonization. To enable flexible interaction between user and harmonization, we introduce interactive harmonization, a new setting where the harmonization is performed with respect to a selected region in the reference image instead of the entire background. A new flexible framework that allows users to pick certain regions of the background image and use it to guide the harmonization is proposed. Inspired by professional portrait harmonization users, we also introduce a new luminance matching loss to optimally match the color/luminance conditions between the composite foreground and select reference region. This framework provides more control to the image harmonization pipeline achieving visually pleasing portrait edits. Furthermore, we also introduce a new dataset carefully curated for validating portrait harmonization. Extensive experiments on both synthetic and real-world datasets show that the proposed approach is efficient and robust compared to previous harmonization baselines, especially for portraits.

****

Self-Distillation for Further Pre-training of Transformers
Seanie Lee,Minki Kang,Juho Lee,Sung Ju Hwang,Kenji Kawaguchi
Pre-training a large transformer model on a massive amount of unlabeled data and fine-tuning it on labeled datasets for diverse downstream tasks has proven to be a successful strategy, for a variety of vision and natural language processing tasks. However, direct fine-tuning of the pre-trained model may be suboptimal if there exist large discrepancies across data domains for pre-training and fine-tuning. To tackle this issue, several previous studies have proposed further pre-training strategies, where we continue to pre-train the model on the target unlabeled dataset before fine-tuning. However, all of them solely focus on language models and we empirically find that a Vision Transformer is vulnerable to overfitting as we continue to pretrain the model on target unlabeled data. In order to tackle this limitation, we propose self-distillation as a regularization for a further pre-training stage. Specifically, we first further pre-train the initial pre-trained model on the target unlabeled data and then consider it as a teacher for self-distillation. Then we take the same initial pre-trained model as a student and enforce its hidden representations to be close to those of the teacher while optimizing the student with a masked auto-encoding objective. We empirically validate the efficacy of self-distillation on a variety of benchmark datasets for image and text classification tasks. Experimentally, we show that our proposed method outperforms all the relevant baselines. Theoretically, we analyze the proposed method with a simplified model to understand how self-distillation for further pre-training can potentially help improve the performance of the downstream tasks.

****

Iterative Relaxing Gradient Projection for Continual Learning
Zeyuan Yang,Zonghan Yang,Peng Li,Yang Liu
A critical capability for intelligent systems is to continually learn given a sequence of tasks. An ideal continual learner should be able to avoid catastrophic forgetting and effectively leverage past learned experiences to master new knowledge. Among different continual learning algorithms, gradient projection approaches impose hard constraints on the optimization space for new tasks to minimize task interference, yet hinder forward knowledge transfer at the same time. Recent methods use expansion-based techniques to relax the constraints, but a growing network can be computationally expensive. Therefore, it remains a challenge whether we can improve forward knowledge transfer for gradient projection approaches \textit{using a fixed network architecture}. In this work, we propose the Iterative Relaxing Gradient Projection (IRGP) framework. The basic idea is to iteratively search for the parameter subspaces most related to the current task and relax these parameters, then reuse the frozen spaces to facilitate forward knowledge transfer while consolidating previous knowledge. Our framework requires neither memory buffers nor extra parameters. Extensive experiments have demonstrated the superiority of our framework over several strong baselines. We also provide theoretical guarantees for our iterative relaxing strategies.

****

Adversarial Counterfactual Environment Model Learning
Xiong-Hui Chen,Yang Yu,Zhengmao Zhu,ZhiHua Yu,Chen Zhenjun,Chenghe Wang,Yinan Wu,Hongqiu Wu,Rong-Jun Qin,Ruijin Ding,Huang Fangsheng
A good model for action-effect prediction, i.e., the environment model, is essential for sample-efficient policy learning, in which the agent can take numerous free trials to find good policies. Currently, the model is commonly learned by fitting historical transition data through empirical risk minimization (ERM). However, we discover that simple data fitting can lead to a model that will be totally wrong in guiding policy learning due to the selection bias in offline dataset collection. In this work, we introduce weighted empirical risk minimization (WERM) to handle this problem in model learning. A typical WERM method utilizes inverse propensity scores to re-weight the training data to approximate the target distribution. However, during the policy training, the data distributions of the candidate policies can be various and unknown. Thus, we propose an adversarial weighted empirical risk minimization (AWRM) objective that learns the model wi

th respect to the worst case of the target distributions. We implement AWRM in a sequential decision structure, resulting in the GALILEO model learning algorithm. We also discover that GALILEO is closely related to adversarial model learning, explaining the empirical effectiveness of the latter. We apply GALILEO in synthetic tasks and verify that GALILEO makes accurate predictions on counterfactual data. We finally applied GALILEO in real-world offline policy learning tasks and found that GALILEO significantly improves policy performance in real-world testing.

****

Admeta: A Novel Double Exponential Moving Average to Adaptive and Non-adaptive Momentum Optimizers with Bidirectional Looking

Yineng Chen,Zuchao Li,Lefei Zhang,Bo Du,hai zhao

Optimizer is an essential component for the success of deep learning, which guides the neural network to update the parameters according to the loss on the training set. SGD and Adam are two classical and effective optimizers on which researchers have proposed many variants, such as SGDM and RAdam. In this paper, we innovatively combine the backward-looking and forward-looking aspects of the optimizer algorithm and propose a novel \textsc{Admeta} (\textbf{A} \textbf{D}ouble exponential \textbf{M}oving averag\textbf{E} \textbf{T}o \textbf{A}daptive and non-adaptive momentum) optimizer framework. For backward-looking part, we propose a DEMA variant scheme, which is motivated by a metric in the stock market, to replace the common exponential moving average scheme. While in the forward-looking part, we present a dynamic lookahead strategy which asymptotically approaching a set value, maintaining its speed at early stage and high convergence performance at final stage. Based on this idea, we provide two optimizer implementations, \textsc{AdmetaR} and \textsc{AdmetaS}, the former based on RAdam and the latter based on SGDM. Through extensive experiments on diverse tasks, we find that the proposed \textsc{Admeta} optimizer outperforms our base optimizers and shows advantages over recently proposed competitive optimizers. We also provide theoretical proof of these two algorithms, which verifies the convergence of our proposed \textsc{Admeta}.

****

Learning from Interval-valued Data

Guangzhi Ma,Jie Lu,Zhen Fang,Feng Liu,Guangquan Zhang

The classification problem concerning crisp-valued data has been well resolved. However, interval-valued data, where all of the observations’ features are described by intervals, is also a common type of data in real-world scenarios. For example, the data extracted by many measuring devices are not exact numbers but intervals. In this paper, we focus on a highly challenging problem called learning from interval-valued data (LIND), where we aim to learn a classifier with high performance on interval-valued observations. First, we obtain the estimation error bound of the LIND problem based on Rademacher complexity. Then, we give the theoretical analysis to show the strengths of multi-view learning on classification problems, which inspires us to construct a new framework called multi-view interval information extraction (Mv-IIE) approach for improving classification accuracy on interval-valued data. The experiment comparisons with several baselines on both synthetic and real-world datasets illustrate the superiority of the proposed framework in handling interval-valued data. Moreover, we describe an application of the Mv-IIE framework that we can prevent data privacy leakage by transforming crisp-valued (raw) data into interval-valued data.

****

Efficient Hyperdimensional Computing

Zhanglu Yan,Shida Wang,Kaiwen Tang,Weng-Fai Wong

Hyperdimensional computing (HDC) uses binary vectors of high dimensions to perform classification. Due to its simplicity and massive parallelism, HDC can be highly energy-efficient and well-suited for resource-constrained platforms. However, in trading off orthogonality with efficiency, hypervectors may use tens of thousands of dimensions. In this paper, we will examine the necessity for such high dimensions. In particular, we give a detailed theoretical analysis of the relationship among dimensions of hypervectors, accuracy, and orthogonality. The main conclusion of this study is that a much lower dimension, typically less than 100, can also achieve similar or even higher detecting accuracy compared with other state-of-the-art HDC models. Based on this insight, we propose a suite of novel techniques to build HDC models that use binary hypervectors of dimensions that are orders of magnitude smaller than those found in the state-of-the-art HDC models, yet yield equivalent or even improved accuracy and efficiency. For image classification, we achieved an HDC accuracy of 96.88\% with a dimension of only 32 on the MNIST dataset. We further explore our methods on more complex datasets like CIFAR-10 and show the limits of HDC computing.

****

Contextual Convolutional Networks

Shuxian Liang,Xu Shen,Tongliang Liu,Xian-Sheng Hua

This paper presents a new Convolutional Neural Network, named Contextual Convolutional Network, that capably serves as a general-purpose backbone for visual recognition. Most existing convolutional backbones follow the representation-to-classification paradigm, where representations of the input are firstly generated by category-agnostic convolutional operations, and then fed into classifiers for specific perceptual tasks (e.g., classification and segmentation). In this paper, we deviate from this classic paradigm and propose to augment potential category memberships as contextual priors in the convolution for contextualized representation learning. Specifically, top-k likely classes from the preceding stage are encoded as a contextual prior vector. Based on this vector and the preceding features, offsets for spatial sampling locations and kernel weights are generated to modulate the convolution operations. The new convolutions can readily replace their plain counterparts in existing CNNs and can be easily trained end-to-end by standard back-propagation without additional supervision. The qualities of Contextual Convolutional Networks make it compatible with a broad range of vision tasks and boost the state-of-the-art architecture ConvNeXt-Tiny by 1.8% on top-1 accuracy of ImageNet classification. The superiority of the proposed model reveals the potential of contextualized representation learning for vision tasks. Code is available at: \url{https://github.com/liang4sx/contextual_cnn}.

****

Factor Learning Portfolio Optimization Informed by Continuous-Time Finance Models

Sinong Geng,houssam nassif,Zhaobin Kuang,Anders Max Reppen,K. Ronnie Sircar

We study financial portfolio optimization in the presence of unknown and uncontrolled system variables referred to as stochastic factors. Existing work falls into two distinct categories: (i) reinforcement learning employs end-to-end policy learning with flexible factor representation, but does not precisely model the dynamics of asset prices or factors; (ii) continuous-time finance methods, in contrast, take advantage of explicitly modeled dynamics but pre-specify, rather than learn, factor representation. We propose FaLPO (factor learning portfolio optimization), a framework that interpolates between these two approaches. Specifically, FaLPO hinges on deep policy gradient to learn a performant investment policy that takes advantage of flexible representation for stochastic factors. Meanwhile, FaLPO also incorporates continuous-time finance models when modeling the dynamics. It uses the optimal policy functional form derived from such models and optimizes an objective that combines policy learning and model calibration. We prove the convergence of FaLPO, and provide performance guarantees via a finite-sample bound. On both synthetic and real-world portfolio optimization tasks, we observe that FaLPO outperforms five leading methods. Finally, we show that FaLPO can be extended to other decision-making problems with stochastic factors.

****

An Incremental Learning Approach for Sustainable Regional Isolation and Integration

Ziyi Wu,Hongge Yao,Jiaojiao Ma

Humans are capable of acquiring new knowledge on a constant basis, while integrating and optimizing old knowledge without forgetting them. This is mainly attributed to the human brain’s ability of partitioned learning and memory replay. In this paper, we simulate this ability and propose an incremental learning network of Sustainable Regional Isolation and Integration (SRII). SRII consists of two phases, regional isolation and regional integration, which are iterated to achieve continuous incremental class learning. Regional isolation isolates new learning processes to avoid interfering with existing knowledge, while regional integration introduces knowledge distillation and margin loss regularization term, knowledge distillation to transfer replay knowledge for alleviating catastrophic forgetting, margin loss regularization term to clarify the boundaries of new and old knowledge for alleviating recency bias. Experimental results on the CIFAR100 and miniImageNet datasets demonstrate that SRII outperforms the state-of-the-arts to avoid catastrophic forgetting. In all 5-stage and 10-stage incremental settings, SRII outperforms the baseline and achieves at least $5.27\%+$ average accuracy improvement. Our source code is available at https://github.com/Wuziyi123/SRII.

****

GraphPNAS: Learning Distribution of Good Neural Architectures via Deep Graph Generative Models

Muchen Li,Jeffrey Yunfan Liu,Leonid Sigal,Renjie Liao

Neural architectures can be naturally viewed as computational graphs. Motivated by this perspective, we, in this paper, study neural architecture search (NAS) through the lens of learning random graph models. In contrast to existing NAS methods which largely focus on searching for a single best architecture, i.e, point estimation, we propose GraphPNAS a deep graph generative model that learns a distribution of well-performing architectures. Relying on graph neural networks (GNNs), our GraphPNAS can better capture topologies of good neural architectures and relations between operators therein. Moreover, our graph generator leads to a learnable probabilistic search method that is more flexible and efficient than the commonly used RNN generator and random search methods. Finally, we learn our generator via an efficient reinforcement learning formulation for NAS. To assess the effectiveness of our GraphPNAS, we conduct extensive experiments on three search spaces, including the challenging RandWire on TinyImageNet, ENAS on CIFAR10, and NAS-Bench-101. The complexity of RandWire is significantly larger than other search spaces in the literature. We show that our proposed graph generator consistently outperforms RNN-based one and achieves better or comparable performances than state-of-the-art NAS methods.

****

Private GANs, Revisited

Alex Bie,Gautam Kamath,Guojun Zhang

We show that with improved training, the standard approach for differentially private GANs -- updating the discriminator with noisy gradients -- achieves or competes with state-of-the-art results for private image synthesis. Existing instantiations of this approach neglect to consider how adding noise only to discriminator updates disrupts the careful balance between generator and discriminator necessary for successful GAN training. We show that a simple fix restores parity: taking more discriminator steps between generator steps. Finally, with the goal of restoring parity between generator and discriminator, we experiment with further modifications to improve discriminator training and see further improvements. For MNIST at $\eps=10$, our private GANs improve the record FID from 48.4 to 13.0, as well as downstream classifier accuracy from 83.2\% to 95.0\%.

****

Hidden Poison: Machine unlearning enables camouflaged poisoning attacks

Jimmy Z. Di,Jack Douglas,Jayadev Acharya,Gautam Kamath,Ayush Sekhari

We introduce camouflaged data poisoning attacks, a new attack vector that arises in the context of machine unlearning and other settings when model retraining may be induced. An adversary first adds a few carefully crafted points to the training dataset such that the impact on the model's predictions is minimal. The adversary subsequently triggers a request to remove a subset of the introduced points at which point the attack is unleashed and the model's predictions are negatively affected. In particular, we consider clean-label targeted attacks (in which the goal is to cause the model to misclassify a specific test point) on datase ts including CIFAR-10, Imagenette, and Imagewoof. This attack is realized by constructing camouflage datapoints that mask the effect of a poisoned dataset.

****

Statistical Inference for Fisher Market Equilibrium

Luofeng Liao,Yuan Gao,Christian Kroer

Statistical inference under market equilibrium effects has attracted increasing attention recently. In this paper we focus on the specific case of linear Fisher markets. They have been widely use in fair resource allocation of food/blood donations and budget management in large-scale Internet ad auctions. In resource allocation, it is crucial to quantify the variability of the resource received by the agents (such as blood banks and food banks) in addition to fairness and efficiency properties of the systems. For ad auction markets, it is important to establish statistical properties of the platform's revenues in addition to their expected values. To this end, we propose a statistical framework based on the concept of infinite-dimensional Fisher markets. In our framework, we observe a market formed by a finite number of items sampled from an underlying distribution (the ``observed market'') and aim to infer several important equilibrium quantities of the underlying long-run market. These equilibrium quantities include individual utilities, social welfare, and pacing multipliers. Through the lens of sample average approximation (SSA), we derive a collection of statistical results and show that the observed market provides useful statistical information of the long-run market. In other words, the equilibrium quantities of the observed market converge to the true ones of the long-run market with strong statistical guarantees. These include consistency, finite sample bounds, asymptotics, and confidence. As an extension, we discuss revenue inference in quasilinear Fisher markets.

****

Auxiliary task discovery through generate and test

Banafsheh Rafiee,Sina Ghiassian,Jun Jin,Richard S. Sutton,Jun Luo,Adam White

In this paper, we explore an approach to auxiliary task discovery in reinforcement learning based on ideas from representation learning. Auxiliary tasks tend to improve data efficiency by forcing the agent to learn auxiliary prediction and control objectives in addition to the main task of maximizing reward, and thus producing better representations. Typically these tasks are designed by people. Meta-learning offers a promising avenue for automatic task discovery; however, these methods are computationally expensive and challenging to tune in practice. In this paper, we explore a complementary approach to the auxiliary task discovery: continually generating new auxiliary tasks and preserving only those with high utility. We also introduce a new measure of auxiliary tasks' usefulness based on how useful the features induced by them are for the main task. Our discovery algorithm significantly outperforms random tasks, hand-designed tasks, and learning without auxiliary tasks across a suite of environments.

****

MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition

Ziqi Gao,Jianguo Chen,Xin Liu,Yuntao Wang,Yuanchun Shi

Multimodal sensors (e.g., visual, non-visual, and wearable) provide complementary information to develop robust perception systems for recognizing activities. However, most existing algorithms use dense sampling and heterogeneous sub-network to extract unimodal features and fuse them at the end of their framework, which causes data redundancy, lack of multimodal complementary information and high computational cost. In this paper, we propose a new novel multi-modal neural architecture based on RGB and IMU wearable sensors (e.g., accelerometer, gyroscope) for human activity recognition called Multimodal Temporal Segment Attention Network (MMTSA). MMTSA first employs a multimodal data isomorphism mechanism based on Gramian Angular Field (GAF) and then applies a novel multimodal sparse sampling method to reduce redundancy. Moreover, we propose an inter-segment attention module in MMTSA to fuse multimodal features effectively and efficiently. We demonstrate the importance of imu data imaging and attention mechanism in human activity recognition by rigours evaluation on three public datasets, and achieved s uperior improvements (11.13% on the MMAct dataset) than the previous state-of-the-art methods.

****

Scenario-based Question Answering with Interacting Contextual Properties
Haitian Sun,William W. Cohen,Ruslan Salakhutdinov

In the scenario-based Question Answering (QA) task, models are asked to find answers that are appropriate to the user scenarios associated with the question and identify information that is missing from the scenarios but is necessary for the answers to hold. Scenarios commonly include multiple properties of users, such as age, employment status, and income level for the question “How much can I claim from this benefit”. The properties relevant to a potential answer are given in a document, which will state conditions necessary for the answer to hold. Documents also may specify how conditions interact with each other, e.g. with text like “one of the conditions below must apply”. Although understanding the relationship between conditions is crucial for solving this challenging QA task, limited work has been done so far in modeling this. In this paper, we propose the T-Reasoner model, which solves this problem with three jointly learned modules: an entailment module which checks whether a condition has been satisfied by the scenario, a decoding module which locates eligible answers from documents, and a reasoning module which infers the relationship between conditions and performs a reasoning step to determine the logically consistent answers and identify missing conditions. T-Reasoner outperforms strong baselines on a synthetic scenario-based QA dataset and achieves a new state-of-the-art on two scenario-based QA benchmarks, outperforming the prior best models by 3-10 points.

****

Easy Differentially Private Linear Regression
Kareem Amin,Matthew Joseph,Mónica Ribero,Sergei Vassilvitskii

Linear regression is a fundamental tool for statistical analysis. This has motivated the development of linear regression methods that also satisfy differential privacy and thus guarantee that the learned model reveals little about any one data point used to construct it. However, existing differentially private solutions assume that the end user can easily specify good data bounds and hyperparameters. Both present significant practical obstacles. In this paper, we study an algorithm which uses the exponential mechanism to select a model with high Tukey depth from a collection of non-private regression models. Given n samples of d-dimensional data used to train m models, we construct an efficient analogue using an approximate Tukey depth that runs in time $O(d^2n + dm\log(m))$. We find that this algorithm obtains strong empirical performance in the data-rich setting with no data bounds or hyperparameter selection required.

****

PointDP: Diffusion-driven Purification against 3D Adversarial Point Clouds
Jiachen Sun,Jiongxiao Wang,Weili Nie,Zhiding Yu,Zhuoqing Mao,Chaowei Xiao

3D Point cloud is a critical data representation in many real-world applications, such as autonomous driving, robotics, and medical imaging. Although the success of deep learning further accelerates the adoption of 3D point clouds in the physical world, deep learning is notoriously vulnerable to adversarial attacks. Various defense solutions have been proposed to build robust models against adversarial attacks. In this work, we identify that the state-of-the-art empirical defense, adversarial training, has a major limitation in 3D point cloud models due to gradient obfuscation, resulting in significant degradation of robustness against strong attacks. To bridge the gap, we propose PointDP, a purification strategy that leverages diffusion models to defend against 3D adversarial attacks. Since PointDP does not rely on predefined adversarial examples for training, it can defend against diverse threats. We extensively evaluate PointDP on six representative 3D point cloud architectures and leverage sixteen strong and adaptive attacks to demonstrate its lower-bound robustness. Our evaluation shows that PointDP achieves significantly better (i.e., 12.6\%-40.3\%) adversarial robustness than state-of-the-art methods under strong attacks bounded by different ℓ_p norms.

****

Visual Recognition with Deep Nearest Centroids

Wenguan Wang,Cheng Han,Tianfei Zhou,Dongfang Liu

We devise deep nearest centroids (DNC), a conceptually elegant yet surprisingly effective network for large-scale visual recognition, by revisiting Nearest Centroids, one of the most classic and simple classifiers. Current deep models learn the classifier in a fully parametric manner, ignoring the latent data structure and lacking simplicity and explainability. DNC instead conducts nonparametric, case-based reasoning; it utilizes sub-centroids of training samples to describe class distributions and clearly explains the classification as the proximity of test data and the class sub-centroids in the feature space. Due to the distance-based nature, the network output dimensionality is flexible, and all the learnable parameters are only for data embedding. That means all the knowledge learnt for ImageNet classification can be completely transferred for pixel recognition learning, under the ‘pre-training and fine-tuning’ paradigm. Apart from its nested simplicity and intuitive decision-making mechanism, DNC can even possess ad-hoc explainability when the sub-centroids are selected as actual training images that humans can view and inspect. Compared with parametric counterparts, DNC performs better on image classification (CIFAR-10, ImageNet) and greatly boots pixel recognition (ADE20K, Cityscapes), with improved transparency and fewer learnable parameters, using various network architectures (ResNet, Swin) and segmentation models (FCN, DeepLabV3, Swin). We feel this work brings fundamental insights into related fields. Our code is available at https://github.com/ChengHan111/DNC.

****

Closing the Gap Between SVRG and TD-SVRG with Gradient Splitting

Arsenii Mustafin,Ioannis Paschalidis,Alex Olshevsky

Temporal difference (TD) learning is a simple algorithm for policy evaluation in reinforcement learning. The performance of TD learning is affected by high variance and it can be naturally enhanced with variance reduction techniques, such as the Stochastic Variance Reduced Gradient (SVRG) method. Recently, multiple works have sought to fuse TD learning with SVRG to obtain a policy evaluation method with a linear rate of convergence. However, the resulting convergence rate is significantly weaker than what is achieved by SVRG in the setting of convex optimization. In this work we utilize a recent interpretation of TD-learning as the splitting of the gradient of an appropriately chosen function, thus simplifying the algorithm and fusing TD with SVRG. We prove a linear convergence bound that is identical to the convergence bound available for SVRG in the convex setting.

****

Rethinking Backdoor Data Poisoning Attacks in the Context of Semi-Supervised Learning

Marissa Catherine Connor,Vincent Emanuele

Semi-supervised learning methods can train high-accuracy machine learning models with a fraction of the labeled training samples required for traditional supervised learning. Such methods do not typically involve close review of the unlabeled training samples, making them tempting targets for data poisoning attacks. In this paper we investigate the vulnerabilities of semi-supervised learning methods to backdoor data poisoning attacks on the unlabeled samples. We show that a simple poisoning attack using adversarially perturbed samples is highly effective - achieving an average attack success rate of 93.6%. We introduce a generalized attack framework targeting semi-supervised learning methods to better understand and exploit their limitations and to motivate future defense strategies.

****

LPT: Long-tailed Prompt Tuning for Image Classification

Bowen Dong,Pan Zhou,Shuicheng Yan,Wangmeng Zuo

For long-tailed classification tasks, most works often pretrain a big model on a large-scale (unlabeled) dataset, and then fine-tune the whole pretrained model for adapting to long-tailed data. Though promising, fine-tuning the whole pret rained model tends to suffer from high cost in computation and deployment of different models for different tasks, as well as weakened generalization capability for overfitting to certain features of long-tailed data. To alleviate these issues, we propose an effective Long-tailed Prompt Tuning (LPT) method for long-tailed classification tasks. LPT introduces several trainable prompts into a frozen pretrained model to adapt it to long-tailed data. For better effectiveness, we divide prompts into two groups: 1) a shared prompt for the whole long-tailed dataset to learn general features and to adapt a pretrained model into the target long-tailed domain; and 2) group-specific prompts to gather group-specific features for the samples which have similar features and also to empower the pretrained model with fine-grained discrimination ability. Then we design a two-phase training paradigm to learn these prompts. In the first phase, we train the shared prompt via conventional supervised prompt tuning to adapt a pretrained model to the desired long-tailed domain. In the second phase, we use the learnt shared prompt as query to select a small best matched set for a group of similar samples from the group-specific prompt set to dig the common features of these similar samples, and then optimize these prompts with a dual sampling strategy and the asymmetric Gaussian Clouded Logit loss. By only fine-tuning a few prompts while fixing the pretrained model, LPT can reduce training cost and deployment cost by storing a few prompts, and enjoys a strong generalization ability of the pretrained model. Experiments show that on various long-tailed benchmarks, with only $\sim$1.1\% extra trainable parameters, LPT achieves comparable or higher performance than previous whole model fine-tuning methods, and is more robust to domain-shift.

****

Interpretable Out-of-Distribution Detection using Pattern Identification

Romain Xu-Darme,Julien Girard-Satabin,Darryl Hond,Gabriele Incorvaia,Zakaria Chihani

Out-of-distribution (OoD) detection for data-based programs is a goal of paramount importance. Common approaches in the literature tend to train binary classifiers requiring inside-of-distribution (IoD) and OoD validation samples, and/or implement confidence metrics that are often abstract and therefore difficult to interpret. In this work, we propose to use the PARTICUL pattern identification algorithm in order to build more interpretable and robust OoD detectors for visual classifiers. Crucially, this approach does not require retraining the classifier and is tuned directly to the IoD dataset, making it applicable to domains where OoD does not have a clear definition. Moreover, pattern identification allows us to provide images from the IoD dataset as reference points to better explain our confidence scores. We illustrate the generalization abilities of our approach through an extensive benchmark across four datasets and two definitions of OoD. Our experiments show that the robustness of all metrics under test does not solely depend on the nature of the IoD dataset or the OoD definition, but also on the architecture of the classifier, which stresses the need for thorough experimentations for future work in OoD detection.

****

TopoZero: Digging into Topology Alignment on Zero-Shot Learning

Yang Liu,Fei Wang,Jiankang Deng,Chen WeiTao,Lei Shang,Baigui Sun,Xuansong Xie

Common space learning, associating semantic and visual domains in a common latent space, is essential to transfer knowledge from seen classes to unseen ones on Zero-Shot Learning (ZSL) realm. Existing methods for common space learning rely heavily on structure alignment due to the heterogeneous nature between semantic and visual domains, but the existing design is sub-optimal. In this paper, we utilize persistent homology to investigate geometry structure alignment, and observe two following issues: (i) The sampled mini-batch data points present a distinct structure gap compared to global data points, thus the learned structure alignment space inevitably neglects abundant and accurate global structure information. (ii) The latent visual and semantic space fail to preserve multiple

dimensional geometry structure, especially high dimensional structure information.
To address the first issue, we propose a Topology-guided Sampling Strategy (TGSS) to mitigate the gap between sampled and global data points. Both theoretical analyses and empirical results guarantee the effectiveness of the TGSS.
To solve the second issue, we introduce a Topology Alignment Module (TAM) to preserve multi-dimensional geometry structure in latent visual and semantic space, respectively. The proposed method is dubbed TopoZero. Empirically, our TopoZero achieves superior performance on three authoritative ZSL benchmark datasets.

****

DamoFD: Digging into Backbone Design on Face Detection

Yang Liu,Jiankang Deng,Fei Wang,Lei Shang,Xuansong Xie,Baigui Sun

Face detection (FD) has achieved remarkable success over the past few years, yet, these leaps often arrive when consuming enormous computation costs. Moreover, when considering a realistic situation, i.e., building a lightweight face detector under a computation-scarce scenario, such heavy computation cost limits the application of the face detector. To remedy this, several pioneering works design tiny face detectors through off-the-shelf neural architecture search (NAS) technologies, which are usually applied to the classification task. Thus, the searched architectures are sub-optimal for the face detection task since some design criteria between detection and classification task are different. As a representative, the face detection backbone design needs to guarantee the stage-level detection ability while it is not required for the classification backbone. Furthermore, the detection backbone consumes a vast body of inference budgets in the whole detection framework.
Considering the intrinsic design requirement and the virtual importance role of the face detection backbone, we thus ask a critical question: How to employ NAS to search FD-friendly backbone architecture? To cope with this question, we propose a distribution-dependent stage-aware ranking score (DDSAR-Score) to explicitly characterize the stage-level expressivity and identify the individual importance of each stage, thus satisfying the aforementioned design criterion of the FD backbone. Based on our proposed DDSAR-Score, we conduct comprehensive experiments on the challenging Wider Face benchmark dataset and achieve dominant performance across a wide range of compute regimes. In particular, compared to the tiniest face detector SCRFD-0.5GF, our method is +2.5 % better in Average Precision (AP) score when using the same amount of FLOPs. The code is avaliable at https://github.com/ly19965/EasyFace/tree/master/face_project/face_detection/DamoFD.

****

Towards Stable Test-time Adaptation in Dynamic Wild World

Shuaicheng Niu,Jiaxiang Wu,Yifan Zhang,Zhiquan Wen,Yaofo Chen,Peilin Zhao,Mingkui Tan

Test-time adaptation (TTA) has shown to be effective at tackling distribution shifts between training and testing data by adapting a given model on test samples. However, the online model updating of TTA may be unstable and this is often a key obstacle preventing existing TTA methods from being deployed in the real world. Specifically, TTA may fail to improve or even harm the model performance when test data have: 1) mixed distribution shifts, 2) small batch sizes, and 3) online imbalanced label distribution shifts, which are quite common in practice. In this paper, we investigate the unstable reasons and find that the batch norm layer is a crucial factor hindering TTA stability. Conversely, TTA can perform more stably with batch-agnostic norm layers, i.e., group or layer norm. However, we observe that TTA with group and layer norms does not always succeed and still suffers many failure cases. By digging into the failure cases, we find that certain noisy test samples with large gradients may disturb the model adaption and result in collapsed trivial solutions, i.e., assigning the same class label for all samples. To address the above collapse issue, we propose a sharpness-aware and reliable entropy minimization method, called SAR, for further stabilizing TTA from two aspects: 1) remove partial noisy samples with large gradients, 2) encourage model weights to go to a flat minimum so that the model is robust to the remaining noisy samples. Promising results demonstrate that SAR performs more stably than prior methods and is computationally efficient under the above wild test scenarios.

****

Sorted eigenvalue comparison d_{Eig}: A simple alternative to d_{FID}

Jiqing Wu,Viktor Koelzer

For $i = 1, 2$, let $\mathbf{S}_i$ be the sample covariance of $\mathbf{Z}_i$ with n_i p-dimensional vectors. First, we theoretically justify an improved Fréchet Inception Distance (d_{FID}) algorithm that replaces np.trace(sqrtm($\mathbf{S}_1 \mathbf{S}_2$)) with np.sqrt(eigvals($\mathbf{S}_1 \mathbf{S}_2$)).sum(). With the appearance of unsorted eigenvalues in the improved d_{FID}, we are then motivated to propose sorted eigenvalue comparison (d_{Eig}) as a simple alternative: $d_{\mathsf{Eig}}(\mathbf{S}_1, \mathbf{S}_2)^2=\sum_{j=1}^p (\sqrt{\lambda_j^1} - \sqrt{\lambda_j^2})^2$, and λ_j^i is the j-th largest eigenvalue of $\mathbf{S}_i$. Second, we present two main takeaways for the improved d_{FID} and proposed d_{Eig} . (i) d_{FID}: The error bound for computing non-negative eigenvalues of diagonalizable $\mathbf S_1 \mathbf S_2$ is reduced to $\mathcal{O}(\varepsilon) \|\mathbf S_1 \| \|\mathbf S_1 \mathbf S_2 \|$, along with reducing the run time by $\sim25\%$. (ii) d_{Eig}: The error bound for computing non-negative eigenvalues of sample covariance $\mathbf S_i$ is further tightened to $\mathcal{O}(\varepsilon) \|\mathbf S_i \|$, with reducing $\sim90\%$ run time. Taking a statistical viewpoint (random matrix theory) on S_i, we illustrate the asymptotic stability of its largest eigenvalues, i.e., rigidity estimates of $\mathcal{O}(n_i^{-\frac{1}{2}+\alpha})$. Last, we discuss limitations and future work for d_{Eig}.

****

Towards Smooth Video Composition

Qihang Zhang,Ceyuan Yang,Yujun Shen,Yinghao Xu,Bolei Zhou

Video generation, with the purpose of producing a sequence of frames, requires synthesizing consistent and persistent dynamic contents over time. This work investigates how to model the temporal relations for composing a video with arbitrary number of frames, from a few to even infinite, using generative adversarial networks (GANs). First, towards composing adjacent frames, we show that the alias-free operation for single image generation, together with adequately pre-learned knowledge, bring a smooth frame transition without harming the per-frame quality. Second, through incorporating a temporal shift module (TSM), which is originally designed for video understanding, into the discriminator, we manage to advance the generator in synthesizing more reasonable dynamics. Third, we develop a novel B-Spline based motion representation to ensure the temporal smoothness, and hence achieve infinite-length video generation, going beyond the frame number used in training. We evaluate our approach on a range of datasets and show substantial improvements over baselines on video generation. Code and models are publicly available at \url{https://genforce.github.io/StyleSV}.

****

Deep Dynamic AutoEncoder for Vision BERT Pretraining

Honghao Chen,Xiangwen Kong,Xiangyu Zhang,Xin Zhao,Kaiqi Huang

Recently, masked image modeling (MIM) has demonstrated promising prospects in se lf-supervised representation learning. However, existing MIM frameworks recover all masked patches equivalently, ignoring that the reconstruction difficulty of different patches can vary sharply due to their diverse distance from visible patches. In this paper, we propose Deep Dynamic AutoEncoder (DDAE), a novel MIM framework that dynamically focuses on patch reconstructions with different degrees of difficulty at different pretraining phases and depths of the model. In addition to raw pixel regression, DDAE performs dynamic feature self-distillation for intermediate layers to learn semantic information. Our methodology provides more locality inductive bias for ViTs, especially in deep layers, which inherently makes up for the absence of local prior for self-attention mechanism. Moreover, our core design deep dynamic supervision can be migrated into existing MIM methods (e.g., MAE, BEiT-v2) seamlessly. The Experimental results demonstrate the effectiveness of our approach. As a tokenizer-free framework, the base-size DDAE can achieve 83.5% top-1 accuracy with only 100 epochs pretraining, surpassing MAE and BEiT pretrained for 800 epochs. For a longer pretraining schedule, DDAE achieves 84.3% top-1 accuracy on Imagenet-1K, and 49.3% mIoU on ADE20K for semantic segmentation.

****

Continuous PDE Dynamics Forecasting with Implicit Neural Representations

Yuan Yin,Matthieu Kirchmeyer,Jean-Yves Franceschi,Alain Rakotomamonjy,patrick gallinari

Effective data-driven PDE forecasting methods often rely on fixed spatial and / or temporal discretizations. This raises limitations in real-world applications like weather prediction where flexible extrapolation at arbitrary spatiotemporal locations is required. We address this problem by introducing a new data-driven approach, DINo, that models a PDE's flow with continuous-time dynamics of spatially continuous functions. This is achieved by embedding spatial observations independently of their discretization via Implicit Neural Representations in a small latent space temporally driven by a learned ODE. This separate and flexible treatment of time and space makes DINo the first data-driven model to combine the following advantages. It extrapolates at arbitrary spatial and temporal locations; it can learn from sparse irregular grids or manifolds; at test time, it generalizes to new grids or resolutions. DINo outperforms alternative neural PDE forecasters in a variety of challenging generalization scenarios on representative PDE systems.

****

Adversarial Collaborative Learning on Non-IID Features

Qinbin Li,Bingsheng He,Dawn Song

Federated Learning (FL) has been a popular approach to enable collaborative learning on multiple parties without exchanging raw data. However, the model performance of FL may degrade a lot due to non-IID data. While many FL algorithms focus on non-IID labels, FL on non-IID features has largely been overlooked. Different from typical FL approaches, the paper proposes a new learning concept called ADCOL (Adversarial Collaborative Learning) for non-IID features. Instead of adopting the widely used model-averaging scheme, ADCOL conducts training in an adversarial way: the server aims to train a discriminator to distinguish the representations of the parties, while the parties aim to generate a common representation distribution. Our experiments on three tasks show that ADCOL achieves better performance than state-of-the-art FL algorithms on non-IID features.

****

DiffMimic: Efficient Motion Mimicking with Differentiable Physics

Jiawei Ren,Cunjun Yu,Siwei Chen,Xiao Ma,Liang Pan,Ziwei Liu

Motion mimicking is a foundational task in physics-based character animation. However, most existing motion mimicking methods are built upon reinforcement learning (RL) and suffer from heavy reward engineering, high variance, and slow convergence with hard explorations. Specifically, they usually take tens of hours or even days of training to mimic a simple motion sequence, resulting in poor scalability. In this work, we leverage differentiable physics simulators (DPS) and propose an efficient motion mimicking method dubbed $\textbf{DiffMimic}$. Our key insight is that DPS casts a complex policy learning task to a much simpler state matching problem. In particular, DPS learns a stable policy by analytical gradients with ground-truth physical priors hence leading to significantly faster and stabler convergence than RL-based methods. Moreover, to escape from local optima, we utilize an \textit{Demonstration Replay} mechanism to enable stable gradient backpropagation in a long horizon. Extensive experiments on standard benchmarks show that DiffMimic has a better sample efficiency and time efficiency than existing methods (e.g., DeepMimic). Notably, DiffMimic allows a physically simulated character to learn back-flip after 10 minutes of training and be able to cycle it after 3 hours of training, while DeepMimic requires about a day of training to cycle back-flip. More importantly, we hope DiffMimic can benefit more differentiable animation systems with techniques like differentiable clothes simulation in future research. Our code is available at https://github.com/diffmimic/diffmimic. Qualitative results can be viewed at https://diffmimic-demo-main-g7h0i8.streamlitapp.com.

****

Towards Inferential Reproducibility of Machine Learning Research

Michael Hagmann,Philipp Meier,Stefan Riezler

Reliability of machine learning evaluation --- the consistency of observed evaluation scores across replicated model training runs --- is affected by several sources of nondeterminism which can be regarded as measurement noise. Current tendencies to remove noise in order to enforce reproducibility of research results neglect inherent nondeterminism at the implementation level and disregard crucial interaction effects between algorithmic noise factors and data properties. This limits the scope of conclusions that can be drawn from such experiments. Instead of removing noise, we propose to incorporate several sources of variance, including their interaction with data properties, into an analysis of significance and reliability of machine learning evaluation, with the aim to draw inferences beyond particular instances of trained models. We show how to use linear mixed effects models (LMEMs) to analyze performance evaluation scores, and to conduct statistical inference with a generalized likelihood ratio test (GLRT). This allows us to incorporate arbitrary sources of noise like meta-parameter variations into statistical significance testing, and to assess performance differences conditional on data properties. Furthermore, a variance component analysis (VCA) enables the analysis of the contribution of noise sources to overall variance and the computation of a reliability coefficient by the ratio of substantial to total variance.

****

Knowledge Distillation based Degradation Estimation for Blind Super-Resolution

Bin Xia,Yulun Zhang,Yitong Wang,Yapeng Tian,Wenming Yang,Radu Timofte,Luc Van Gool

Blind image super-resolution (Blind-SR) aims to recover a high-resolution (HR) image from its corresponding low-resolution (LR) input image with unknown degradations. Most of the existing works design an explicit degradation estimator for each degradation to guide SR. However, it is infeasible to provide concrete labels of multiple degradation combinations (\eg, blur, noise, jpeg compression) to supervise the degradation estimator training. In addition, these special designs for certain degradation, such as blur, impedes the models from being generalized to handle different degradations. To this end, it is necessary to design an implicit degradation estimator that can extract discriminative degradation representation for all degradations without relying on the supervision of degradation ground-truth. In this paper, we propose a Knowledge Distillation based Blind-SR network (KDSR). It consists of a knowledge distillation based implicit degradation estimator network (KD-IDE) and an efficient SR network. To learn the KDSR model, we first train a teacher network: KD-IDE$_{T}$. It takes paired HR and LR patches as inputs and is optimized with the SR network jointly. Then, we further train a student network KD-IDE$_{S}$, which only takes LR images as input and learns to extract the same implicit degradation representation (IDR) as KD-IDE$_{T}$. In addition, to fully use extracted IDR, we design a simple, strong, and efficient IDR based dynamic convolution residual block (IDR-DCRB) to build an SR network. We conduct extensive experiments under classic and real-world degradation se ttings. The results show that KDSR achieves SOTA performance and can generalize to various degradation processes. The source codes and pre-trained models will be released.

****

Very Large Scale Multi-Agent Reinforcement Learning with Graph Attention Mean Field

Qianyue Hao

With recent advances in reinforcement learning, we have witnessed countless successes of intelligent agents in various domains. Especially, multi-agent reinforcement learning (MARL) is suitable for many real-world scenarios and has vast potential applications. However, typical MARL methods can only handle tens of agents, leaving scenarios with up to hundreds or even thousands of agents almost unexplored. There exist two key challenges in scaling up the number of agents: (1) agent-agent interactions are critical in multi-agent systems while the number of interactions grows quadratically with the number of agents, causing great computational complexity and difficulty in strategies-learning; (2) the strengths of interactions vary among agents and over time, making it difficult to precisely model such interactions. In this paper, we propose the Graph Attention Mean Field (GAT-MF) method, where we convert agent-agent interactions into interactions between each agent and a weighted mean field, greatly reducing the computational complexity. We mathematically prove the correctness of this conversion. We design a graph attention mechanism to automatically capture the different and time-varying strengths of interactions, ensuring the ability of our method to precisely model interactions among the agents. We conduct extensive experiments in both manual and real-world scenarios with up to more than 3000 agents, demonstrating that comparing existing MARL methods, our method reaches superior performance and 9.4 times computational efficiency.

****

Graph Contrastive Learning for Skeleton-based Action Recognition

Xiaohu Huang,Hao Zhou,Jian Wang,Haocheng Feng,Junyu Han,Errui Ding,Jingdong Wang,Xinggang Wang,Wenyu Liu,Bin Feng

In the field of skeleton-based action recognition, current top-performing graph convolutional networks (GCNs) exploit intra-sequence context to construct adaptive graphs for feature aggregation. However, we argue that such context is still $\textit{local}$ since the rich cross-sequence relations have not been explicitly investigated. In this paper, we propose a graph contrastive learning framework for skeleton-based action recognition ($\textit{SkeletonGCL}$) to explore the $\textit{global}$ context across all sequences. In specific, SkeletonGCL associates graph learning across sequences by enforcing graphs to be class-discriminative, i.e., intra-class compact and inter-class dispersed, which improves the GCN capacity to distinguish various action patterns. Besides, two memory banks are designed to enrich cross-sequence context from two complementary levels, i.e., instance and semantic levels, enabling graph contrastive learning in multiple context scales. Consequently, SkeletonGCL establishes a new training paradigm, and it can be seamlessly incorporated into current GCNs. Without loss of generality, we combine SkeletonGCL with three GCNs (2S-ACGN, CTR-GCN, and InfoGCN), and achieve consistent improvements on NTU60, NTU120, and NW-UCLA benchmarks.

****

Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation

Jie Yang,Ailing Zeng,Shilong Liu,Feng Li,Ruimao Zhang,Lei Zhang

This paper presents a novel end-to-end framework with Explicit box Detection for multi-person Pose estimation, called ED-Pose, where it unifies the contextual learning between human-level (global) and keypoint-level (local) information. Different from previous one-stage methods, ED-Pose re-considers this task as two explicit box detection processes with a unified representation and regression supervision. First, we introduce a human detection decoder from encoded tokens to extract global features. It can provide a good initialization for the latter keypoint detection, making the training process converge fast. Second, to bring in contextual information near keypoints, we regard pose estimation as a keypoint box detection problem to learn both box positions and contents for each keypoint. A human-to-keypoint detection decoder adopts an interactive learning strategy between human and keypoint features to further enhance global and local feature aggregation. In general, ED-Pose is conceptually simple without post-processing and dense heatmap supervision. It demonstrates its effectiveness and efficiency compared with both two-stage and one-stage methods. Notably, explicit box detection boosts the pose estimation performance by 4.5 AP on COCO and 9.9 AP on CrowdPose. For the first time, as a fully end-to-end framework with a L1 regression loss, ED-Pose surpasses heatmap-based Top-down methods under the same backbone by 1.2 AP on COCO and achieves the state-of-the-art with 76.6 AP on CrowdPose without bells and whistles. Code is available at https://github.com/IDEA-Research/ED-Pose.

****

Expected Perturbation Scores for Adversarial Detection

Shuhai Zhang,Feng Liu,Jiahao Yang,Yifan Yang,Bo Han,Mingkui Tan

Adversarial detection aims to determine whether a given sample is an adversarial one based on the discrepancy between natural and adversarial distributions. Unfortunately, estimating or comparing two data distributions is extremely difficult, especially in the high-dimension space. Recently, the gradients of log probability density (a.k.a., score) w.r.t. samples is used as an alternative statistic to compute. However, we find that the score is sensitive in identifying adversarial samples due to insufficient information with one sample only. In this paper, we propose a new statistic called expected perturbation score (EPS), which is essentially the expected score of a sample after various perturbations. Specifically, to obtain adequate information regarding one sample, we can perturb it by adding various noises to capture its multi-view observations. We theoretically prove that EPS is a proper statistic to compute the discrepancy between two distributions under mild conditions. In practice, we can use a pre-trained diffusion model to estimate EPS for each sample. Last, we propose an EPS-based adversarial detection (EPS-AD) method, in which we develop EPS-based maximum mean discrepancy (MMD) as a metric to measure the discrepancy between the test sample and natural samples. To verify the validity of our proposed method, we also prove that the EPS-based MMD between natural and adversarial samples is larger than that among natural samples. Empirical studies on CIFAR-10 and ImageNet across different network architectures including ResNet, WideResNet, and ViT show the superior adversarial detection performance of EPS-AD compared to existing methods.

****

Look Back When Surprised: Stabilizing Reverse Experience Replay for Neural Approximation

Ramnath Kumar,Dheeraj Mysore Nagaraj

Experience replay-based sampling techniques are essential to several reinforcement learning (RL) algorithms since they aid in convergence by breaking spurious correlations. The most popular techniques, such as uniform experience replay(UER) and prioritized experience replay (PER), seem to suffer from sub-optimal convergence and significant bias error, respectively. To alleviate this, we introduce a new experience replay method for reinforcement learning, called Introspective Experience Replay (IER). IER picks batches corresponding to data points consecutively before the ‘surprising’ points. Our proposed approach is based on the theoretically rigorous reverse experience replay (RER), which can be shown to remove bias in the linear approximation setting but can be sub-optimal with neural approximation. We show empirically that IER is stable with neural function approximation and has a superior performance compared to the state-of-the-art techniques like uniform experience replay (UER), prioritized experience replay(PER), and hindsight experience replay (HER) on the majority of tasks.

****

BQ-NCO: Bisimulation Quotienting for Generalizable Neural Combinatorial Optimization

Darko Drakulic,Sofia Michel,Florian Mai,Arnaud Sors,Jean-Marc Andreoli

Despite the success of Neural Combinatorial Optimization methods for end-to-end heuristic learning, out-of-distribution generalization remains a challenge.
In this paper, we present a novel formulation of combinatorial optimization (CO)

problems as Markov Decision Processes (MDPs) that effectively leverages symmetries of the CO problems to improve out-of-distribution robustness.
Starting from the standard MDP formulation of constructive heuristics, we introduce a generic transformation based on bisimulation quotienting (BQ) in MDPs.
This transformation allows to reduce the state space by accounting for the intrinsic symmetries of the CO problem and facilitates the MDP solving.
We illustrate our approach on the Traveling Salesman and Capacitated Vehicle Routing Problems. We present a BQ reformulation of these problems and introduce a simple attention-based policy network that we train by imitation of (near) optimal solutions for small instances from a single distribution.
We obtain new state-of-the-art generalization results for instances with up to 1000 nodes from synthetic and realistic benchmarks that vary both in size and node distributions.
****

CoGANs: Collaborative Generative Adversarial Networks

Omkar Joglekar,Goren Gordon

In complex creative scenarios, co-creativity by multiple agents offer great advantages. Each agent has a specific skill set and a set of abilities, which is sometimes not enough to perform a general, large and complex task single-handed. These kinds of tasks benefit substantially from collaboration. In deep learning applications, data generation is an example of such a complex, potentially multi-modal task. Previous Generative Adversarial Networks (GANs) focused on using a single generator to generate multi-modal datasets, which is sometimes known to face issues such as mode-collapse and failure to converge. The multi-generator based works such as MGAN, MMGAN, MADGAN and AdaGAN either require training a classifier online, the use of complex mixture models or sequentially adding generators, which is computationally complex.

In this work, we present a simple, novel approach of training collaborative GANs (CoGAN), with multiple generators and a single critic/discriminator, without introducing external complexities such as a classifier model. We show that this method of workload division meets the state-of-the-art quality metrics, and makes GAN training robust. We present a proof-of-concept on the MNIST dataset, which has 10 modes of data. The individual generators learn to generate different digits from the distribution, and together learn to generate the whole distribution. We introduce a new component to the generator loss during GAN training, based on the Total Variation Distance (TVD) and show that it significantly improves stability during training and performance over state-of-the-art single generator GANs.
****

Multiscale Neural Operator: Learning Fast and Grid-independent PDE Solvers

Björn Lütjens,Catherine H. Crawford,Campbell D Watson,Christopher Hill,Dava Newman

Numerical simulations in climate, chemistry, or astrophysics are computationally too expensive for uncertainty quantification or parameter-exploration at high-resolution. Reduced-order or surrogate models are multiple orders of magnitude faster, but traditional surrogates are inflexible or inaccurate and pure machine learning (ML)-based surrogates too data-hungry. We propose a hybrid, flexible surrogate model that exploits known physics for simulating large-scale dynamics and limits learning to the hard-to-model term, which is called parametrization or closure and captures the effect of fine- onto large-scale dynamics. Leveraging neural operators, we are the first to learn grid-independent, non-local, and flexible parametrizations. Our \textit{multiscale neural operator} is motivated by a rich literature in multiscale modeling, has quasilinear runtime complexity, is more accurate or flexible than state-of-the-art parametrizations and demonstrated on the chaotic equation multiscale Lorenz96.
****

Out-of-distribution Detection with Diffusion-based Neighborhood

Luping Liu,Yi Ren,Xize Cheng,Zhou Zhao

Out-of-distribution (OOD) detection is an important task to ensure the reliability and safety of deep learning and the discriminator models outperform others fo r now. However, the feature extraction of such models must compress the data and lose certain information, leaving room for bad cases and malicious attacks. However, despite effectively fitting the data distribution and producing high-quality samples, generative models lack suitable indicator scores to match with discriminator models in the OOD detection tasks. In this paper, we find that these two kinds of models can be combined to solve each other's problems. We introduce diffusion models (DMs), a kind of powerful generative model, into OOD detection and find that the denoising process of DMs also functions as a novel form of asymmetric interpolation. This property establishes a diffusion-based neighborhood for each input data. Then, we perform discriminator-based OOD detection based on the diffusion-based neighborhood instead of isolated data. In this combination, the discriminator models provide detection metrics for generation models and the diffusion-based neighborhood reduces the information loss of feature extraction. According to our experiments on CIFAR10 and CIFAR100, our new methods successfully outperform state-of-the-art methods. Our implementation is put in the supplementary materials.

****

Never Revisit: Continuous Exploration in Multi-Agent Reinforcement Learning
Chenghao Li,Tonghan Wang,Xiaoran Wu,Jun Yang,Qianchuan Zhao,Chongjie Zhang
Recently, intrinsic motivations are wildly used for exploration in multi-agent reinforcement learning. We discover that coming with intrinsic rewards is the issue of revisitation -- the relative values of intrinsic rewards fluctuate, causing a sub-space visited before becomes attractive after a period of exploration to other areas. Consequently, agents risk exploring some sub-spaces repeatedly. In this paper, we formally define the concept of revisitation, based on which we propose an observation-distribution matching approach to detect the appearance of revisitation. To avoid it, we add branches to agents' local Q-networks and the mixing network to separate sub-spaces which have already been revisited. Furthermore, to prevent adding branches excessively, we design intrinsic rewards to reduce the probability of and penalize the occurrence of revisitation. By virtue of these advances, our method achieves superior performance on three challenging Google Research Football (GRF) scenarios with sparse rewards.

****

Do Not Train It: A Linear Neural Architecture Search of Graph Neural Networks
Peng XU,Lin Zhang,Xuanzhou Liu,Jiaqi Sun,Yue Zhao,Haiqin Yang,Bei Yu
Neural architecture search (NAS) for Graph neural networks (GNNs), called NAS-GNNs, has achieved significant performance over manually designed GNN architectures. However, these methods inherit issues from the conventional NAS methods, such as high computational cost and optimization difficulty. More importantly, previous NAS methods have ignored the uniqueness of GNNs, where the non-linearity has limited effect. Based on this, we are the first to theoretically prove that a GNN fixed with random weights can obtain optimal outputs under mild conditions. With the randomly-initialized weights, we can then seek the optimal architecture parameters via the sparse coding objective and derive a novel NAS-GNNs method, namely neural architecture coding (NAC). Consequently, our NAC holds a no-update scheme on GNNs and can efficiently compute in linear time. Empirical evaluations on multiple GNN benchmark datasets demonstrate that our approach leads to state-of-the-art performance, which is up to $200\times$ faster and 18.8% more accurate than the strong baselines.

****

Rethinking the Explanation of Graph Neural Network via Non-parametric Subgraph Matching
Fang Wu,Lirong Wu,Siyuan Li,Dragomir Radev,Wenbing Huang
The great success in graph neural networks (GNNs) provokes the question about explainability: ``Which fraction of the input graph is the most determinant to the prediction?'' However, current approaches usually resort to a black-box to decipher another black-box (i.e., GNN), making it difficult to understand how the explanation is made. Based on the observation that graphs typically share some joint motif patterns, we propose a novel subgraph matching framework named MatchExplainer to explore explanatory subgraphs.

It couples the target graph with other counterpart instances and identifies the most crucial joint substructure by minimizing the node corresponding-based distance between them. After that, an external graph ranking is followed to select the most informative substructure from all subgraph candidates. Thus, MatchExplainer is entirely non-parametric.

Moreover, present graph sampling or node dropping methods usually suffer from the false positive sampling problem. To ameliorate that issue, we take advantage of MatchExplainer to fix the most informative portion of the graph and merely operate graph augmentations on the rest less informative part, which is dubbed as MatchDrop.

We conduct extensive experiments on both synthetic and real-world datasets, showing the effectiveness of our MatchExplainer by outperforming all parametric baselines with large margins. Additional results also demonstrate that our MatchDrop is a general paradigm to be equipped with GNNs for enhanced performance.

****

Spikformer: When Spiking Neural Network Meets Transformer

Zhaokun Zhou,Yuesheng Zhu,Chao He,Yaowei Wang,Shuicheng YAN,Yonghong Tian,Li Yuan

We consider two biologically plausible structures, the Spiking Neural Network (SNN) and the self-attention mechanism. The former offers an energy-efficient and event-driven paradigm for deep learning, while the latter has the ability to capture feature dependencies, enabling Transformer to achieve good performance. It is intuitively promising to explore the marriage between them. In this paper, we consider leveraging both self-attention capability and biological properties of SNNs, and propose a novel Spiking Self Attention (SSA) as well as a powerful framework, named Spiking Transformer (Spikformer). The SSA mechanism in Spikformer models the sparse visual feature by using spike-form Query, Key, and Value without softmax. Since its computation is sparse and avoids multiplication, SSA is efficient and has low computational energy consumption. It is shown that Spikformer with SSA can outperform the state-of-the-art SNNs-like frameworks in image classification on both neuromorphic and static datasets. Spikformer (66.3M parameters) with comparable size to SEW-ResNet-152 (60.2M,69.26%) can achieve 74.81% top1 accuracy on ImageNet using 4 time steps, which is the state-of-the-art in directly trained SNNs models. Code is avaiable at https://github.com/ZK-Zhou/spikformer.

****

DeSCo: Towards Scalable Deep Subgraph Counting

Tianyu Fu,Yu Wang,Zhitao Ying

Subgraph counting is the problem of determining the number of a given query graph in a large targe graph. Despite being a #P problem, subgraph counting is a crucial graph analysis method in domains ranging from biology and social science to risk management and software analysis. However, existing exact counting methods take combinatorially long runtime as target and query sizes increase. Existing approximate heuristic methods and neural approaches fall short in accuracy due to high label dynamic range, limited model expressive power, and inability to predict the distribution of subgraph counts in the target graph. Here we propose DeSCo, a neural deep subgraph counting framework, which aims to accurately predict the count and distribution of query graphs on any given target graph. DeSCo uses canonical partition to divide the large target graph into small neighborhood graphs and predict the canonical count objective on each neighborhood. The proposed partition method avoids missing or double-counting any patterns of the target graph. A novel subgraph-based heterogeneous graph neural network is then used to improve the expressive power. Finally, gossip correction improves counting accuracy via prediction propagation with learnable weights. Compared with state-of-the-art approximate heuristic and neural methods. DeSCo achieves 437x improvement in the mean squared error of count prediction and benefits from the polynomial runtime complexity.

****

On a Built-in Conflict between Deep Learning and Systematic Generalization

Yuanpeng Li
Out-of-distribution or systematic generalization is a desirable property that most deep learning algorithms lack. In this paper, we hypothesize that internal function sharing is one of the reasons to weaken systematic generalization in deep learning for classification tasks. Under equivalent prediction, a model partitions an input space into multiple parts separated by boundaries. The function sharing prefers to reuse boundaries, leading to fewer parts for new outputs, which conflicts with systematic generalization. We show such phenomena in standard deep learning models, such as fully connected, convolutional, residual networks, LSTMs, and (Vision) Transformers. We hope this study provides novel insights and forms a basis for new research directions to improve systematic generalization.
****
Consistent and Truthful Interpretation with Fourier Analysis
Yifan Zhang,Haowei He,Yang Yuan
For many interdisciplinary fields,
ML interpretations need to be consistent with \emph{what-if} scenarios related to the current case, i.e., if one factor changes, how does the model react?
Although the attribution methods are supported by the elegant axiomatic systems, they mainly focus on individual inputs,
and are generally inconsistent.
To support what-if scenarios, we introduce a new objective of consistency based on a notion called truthful interpretation. Towards this objective,
we apply Fourier analysis of Boolean functions to get
consistency guarantees.
Experimental results show that
for neighborhoods with various radii,
our method achieves 2x - 50x lower inconsistency compared with the other methods.
****
D2Match: Leveraging Deep Learning and Degeneracy for Subgraph Matching
Xuanzhou Liu,Lin Zhang,Jiaqi Sun,Yujiu Yang,Haiqin Yang
Subgraph matching is a fundamental building block for many graph-based applications and is challenging due to its high-order combinatorial nature. However, previous methods usually tackle it by combinatorial optimization or representation learning and suffer from exponential computational cost or matching without theoretical guarantees. In this paper, we develop D2Match by leveraging the efficiency of Deep learning and Degeneracy for subgraph matching. More specifically, we prove that subgraph matching can degenerate to subtree matching, and subsequently is equivalent to finding a perfect matching on a bipartite graph. This matching procedure can be implemented by the built-in tree-structured aggregation mechanism on graph neural networks, which yields linear time complexity. Moreover, circle structures, abstracted as {\em supernodes}, and node attributes can be easily incorporated in D2Match to boost the matching. Finally, we conduct extensive experiments to show the superior performance of our D2Match and confirm that our D2Match indeed tries to exploit the subtrees and differs from existing learning-based subgraph matching methods that depend on memorizing the data distribution divergence.
****
Multimodal Analogical Reasoning over Knowledge Graphs
Ningyu Zhang,Lei Li,Xiang Chen,Xiaozhuan Liang,Shumin Deng,Huajun Chen
Analogical reasoning is fundamental to human cognition and holds an important place in various fields. However, previous studies mainly focus on single-modal analogical reasoning and ignore taking advantage of structure knowledge. Notably, the research in cognitive psychology has demonstrated that information from multimodal sources always brings more powerful cognitive transfer than single modality sources. To this end, we introduce the new task of multimodal analogical reasoning over knowledge graphs, which requires multimodal reasoning ability with the help of background knowledge. Specifically, we construct a Multimodal Analogical Reasoning dataSet (MARS) and a multimodal knowledge graph MarKG. We evaluate with multimodal knowledge graph embedding and pre-trained Transformer baselines,

illustrating the potential challenges of the proposed task. We further propose a novel model-agnostic Multimodal analogical reasoning framework with Transformer (MarT) motivated by the structure mapping theory, which can obtain better performance. We hope our work can deliver benefits and inspire future research. Code and datasets are available in https://github.com/zjunlp/MKG_Analogy.

****

GAIN: Enhancing Byzantine Robustness in Federated Learning with Gradient Decomposition

Yuchen Liu,Chen Chen,Lingjuan Lyu,Fangzhao Wu,Tianlei Hu,Sai Wu,Gang Chen

Federated learning provides a privacy-aware learning framework by enabling participants to jointly train models without exposing their private data. However, federated learning has exhibited vulnerabilities to Byzantine attacks, where the adversary aims to destroy the convergence and performance of the global model. Meanwhile, we observe that most existing robust AGgregation Rules (AGRs) fail to stop the aggregated gradient deviating from the optimal gradient (the average of honest gradients) in the non-IID setting. We attribute the reason of the failure of these AGRs to two newly proposed concepts: identification failure and integrity failure. The identification failure mainly comes from the exacerbated curse of dimensionality in the non-IID setting. The integrity failure is a combined result of conservative filtering strategy and gradient heterogeneity. In order to address both failures, we propose GAIN, a gradient decomposition scheme that can help adapt existing robust algorithms to heterogeneous datasets. We theoretically show that integrating exisiting robust AGRs into our GAIN can mitigate the deviation of aggregated gradient, thus improve the performance. Experiments on various real-world datasets verify the efficacy of our proposed GAIN

****

Temporary feature collapse phenomenon in early learning of MLPs

Dongrui Liu,Shaobo Wang,Jie Ren,Kangrui Wang,Sheng Yin,Huiqi Deng,Quanshi Zhang

In this paper, we focus on a typical two-phase phenomenon in the learning of multi-layer perceptrons (MLPs). We discover and explain the reason for the feature collapse phenomenon in the first phase, i.e., the diversity of features over different samples keeps decreasing in the first phase, until samples of different categories share almost the same feature, which hurts the optimization of MLPs. We explain such a phenomenon in terms of the learning dynamics of MLPs. Furthermore, we theoretically analyze the reason why four typical operations can alleviate the feature collapse. The code has been attached with the submission.

****

MECTA: Memory-Economic Continual Test-Time Model Adaptation

Junyuan Hong,Lingjuan Lyu,Jiayu Zhou,Michael Spranger

Continual Test-time Adaptation (CTA) is a promising art to secure accuracy gains in continually-changing environments. The state-of-the-art adaptations improve out-of-distribution model accuracy via computation-efficient online test-time gradient descents but meanwhile cost about times of memory versus the inference, even if only a small portion of parameters are updated. Such high memory consumption of CTA substantially impedes wide applications of advanced CTA on memory-constrained devices. In this paper, we provide a novel solution, dubbed MECTA, to drastically improve the memory efficiency of gradient-based CTA. Our profiling shows that the major memory overhead comes from the intermediate cache for back-propagation, which scales by the batch size, channel, and layer number. Therefore, we propose to reduce batch sizes, adopt an adaptive normalization layer to maintain stable and accurate predictions, and stop the back-propagation caching heuristically. On the other hand, we prune the networks to reduce the computation and memory overheads in optimization and recover the parameters afterward to avoid forgetting. The proposed MECTA is efficient and can be seamlessly plugged into state-of-the-art CTA algorithms at negligible overhead on computation and memory. On three datasets, CIFAR10, CIFAR100, and ImageNet, MECTA improves the accuracy by at least 6% with constrained memory and significantly reduces the memory costs of ResNet50 on ImageNet by at least 70% with comparable accuracy. Our codes can be accessed at https://github.com/SonyAI/MECTA.

****

MocoSFL: enabling cross-client collaborative self-supervised learning

Jingtao Li,Lingjuan Lyu,Daisuke Iso,Chaitali Chakrabarti,Michael Spranger

Existing collaborative self-supervised learning (SSL) schemes are not suitable for cross-client applications because of their expensive computation and large local data requirements. To address these issues, we propose MocoSFL, a collaborative SSL framework based on Split Federated Learning (SFL) and Momentum Contrast (MoCo). In MocoSFL, the large backbone model is split into a small client-side model and a large server-side model, and only the small client-side model is processed locally on the client's local devices. MocoSFL has three key components: (i) vector concatenation which enables the use of small batch size and reduces computation and memory requirements by orders of magnitude; (ii) feature sharing that helps achieve high accuracy regardless of the quality and volume of local data; (iii) frequent synchronization that helps achieve better non-IID performance because of smaller local model divergence. For a 1,000-client case with non-IID data (each client only has data from 2 random classes of CIFAR-10), MocoSFL can achieve over 84% accuracy with ResNet-18 model. Next we present TAResSFL module that significantly improves the resistance to privacy threats and communication overhead with small sacrifice in accuracy for a MocoSFL system. On a Raspberry Pi 4B device, the MocoSFL-based scheme requires less than 1MB of memory and less than 40MB of communication, and consumes less than 5W power. The code is available at https://github.com/SonyAI/MocoSFL.

****

Block-level Stiffness Analysis of Residual Networks

Eliska Kloberdanz,Wei Le

Residual Networks (ResNets) can be interpreted as dynamic systems, which are systems whose state changes over time and can be described with ordinary differential equations (ODEs). Specifically, the dynamic systems interpretation views individual residual blocks as ODEs. The solution to an ODE is an approximation; and therefore contains an error term. If an ODE is stiff it is likely that this error is amplified and becomes dominating in the solution calculations, which negatively affects the accuracy of the approximated solution. Therefore, stiff ODEs are often numerically unstable. In this paper we leverage the dynamic systems interpretation to perform a novel theoretical analysis of ResNets by leveraging findings and tools from numerical analysis of ODEs. Specifically, we perform block level stiffness analysis of ResNets. We find that residual blocks towards the end of ResNet models exhibit increased stiffness and that there is a significant correlation between stiffness and model accuracy and loss. Based on these findings, we propose that ResNets behave as stiff numerically unstable ODEs.

****

Q-Match: Self-Supervised Learning For Tabular Data by Matching Distributions Induced by a Queue

Thomas Mulc,Debidatta Dwibedi

In semi-supervised learning, student-teacher distribution matching has been successful in improving performance of models using unlabeled data in conjunction with few labeled samples. In this paper, we aim to replicate that success in the self-supervised setup where we do not have access to any labeled data during pre-training. We show it is possible to induce the student-teacher distributions without any knowledge of downstream classes by using a queue of embeddings of samples from the unlabeled dataset. We show that Q-Match outperforms previous self-supervised learning techniques on tabular datasets when measuring downstream classification performance. Furthermore, we show that our method is sample efficient, both in terms of labels required for both downstream task training and amount of unlabeled data required for pre-training.

****

Supervised Contrastive Regression

Kaiwen Zha,Peng Cao,Yuzhe Yang,Dina Katabi

Deep regression models typically learn in an end-to-end fashion and do not explicitly try to learn a regression-aware representation. Their representations tend to be fragmented and fail to capture the continuous nature of regression tasks. In this paper, we propose Supervised Contrastive Regression (SupCR), a framewor

k that learns a regression-aware representation by contrasting samples against each other based on their target distance. SupCR is orthogonal to existing regression models, and can be used in combination with such models to improve performance. Extensive experiments using five real-world regression datasets that span computer vision, human-computer interaction, and healthcare show that using SupCR achieves the state-of-the-art performance and consistently improves prior regression baselines on all datasets, tasks, and input modalities. SupCR also improves robustness to data corruptions, resilience to reduced training data, performance on transfer learning, and generalization to unseen targets.
****

SELF-SUPERVISED PRETRAINING FOR DIFFERENTIALLY PRIVATE LEARNING
Arash Asadian,Evan Weidner,Lei Jiang

We demonstrate self-supervised pretraining (SSP) is a scalable solution to deep learning with differential privacy (DP) regardless of the size of available public datasets in image classification. When facing the lack of public datasets, we show the features generated by SSP on only one single image enable a private classifier to obtain a much better utility than the non-learned handcrafted features under the same privacy budget. When a moderate or large size public dataset is available, the features produced by SSP greatly outperform the features trained with labels on various complex private datasets under the same private budget. We also compared multiple DP-enabled training frameworks to train a private classifier on the features generated by SSP.
****

Explainable Artificial Intelligence: Reaping the Fruits of Decision Trees
Ralf Peter Riedel,Aviv Segev

The recent push for explainable artificial intelligence (XAI) has given rise to extensive work toward understanding the inner workings of neural networks. Much of that work, however, has focused on manipulating input data feeding the network to assess their effect on network output. It is shown in this study that XAI can benefit from investigating the network node, the most fundamental unit of neural networks. Whereas studies on XAI have mostly benefited from a focus on manipulating input data, assessing patterns in node weights may prove equally beneficial, if not more significant, especially when realizing that weight values may not be as random as previously thought. A manipulated, a contrived, and a real dataset were used in this study. Datasets were run on convolutional and deep neural network models. Node rank stability was the central construct to investigate neuronal patterns in this study. Rank stability was defined as the number of epochs wherein nodes held their rank in terms of weight value compared to their rank at the last epoch, when the model reached convergence, or stability (defined in this study as accuracy $\geq$ 0.90). Findings indicated that neural networks behaved like a decision tree, in that rank stability increased as weight absolute values increased. Decision tree behavior may assist in more efficient pruning algorithms, which may produce distilled models simpler to explain to technical and non-technical audiences.
****

Interpretability with full complexity by constraining feature information
Kieran A Murphy,Danielle Bassett

Interpretability is a pressing issue for machine learning. Common approaches to interpretable machine learning constrain interactions between features of the input, sacrificing model complexity in order to render more comprehensible the effects of those features on the model's output. We approach interpretability from a new angle: constrain the information about the features without restricting the complexity of the model. We use the Distributed Information Bottleneck to optimally compress each feature so as to maximally preserve information about the output. The learned information allocation, by feature and by feature value, provides rich opportunities for interpretation, particularly in problems with many features and complex feature interactions. The central object of analysis is not a single trained model, but rather a spectrum of models serving as approximations that leverage variable amounts of information about the inputs. Information is allocated to features by their relevance to the output, thereby solving the prob lem of feature selection by constructing a learned continuum of feature inclusion-to-exclusion. The optimal compression of each feature---at every stage of approximation---allows fine-grained inspection of the distinctions among feature values that are most impactful for prediction. We develop a framework for extracting insight from the spectrum of approximate models and demonstrate its utility on a range of tabular datasets.

****

Revisiting Group Robustness: Class-specific Scaling is All You Need

Seonguk Seo,Bohyung Han

Group distributionally robust optimization, which aims to improve robust accuracies such as worst-group or unbiased accuracy, is one of the mainstream algorithms to mitigate spurious correlation and reduce dataset bias. While existing approaches have apparently gained performance in robust accuracy, these improvements mainly come from a trade-off at the expense of average accuracy. To address the challenges, we first propose a simple class-specific scaling strategy to control the trade-off between robust and average accuracies flexibly and efficiently, which is directly applicable to existing debiasing algorithms without additional training; it reveals that a naive ERM baseline matches or even outperforms the recent debiasing approaches by adopting the class-specific scaling. Then, we employ this technique to 1) evaluate the performance of existing algorithms in a comprehensive manner by introducing a novel unified metric that summarizes the trade-off between the two accuracies as a scalar value and 2) develop an instance-wise adaptive scaling technique for overcoming the trade-off and improving the performance even further in terms of both accuracies. Experimental results verify the effectiveness of the proposed frameworks in both tasks.

****

Provable Benefits of Representational Transfer in Reinforcement Learning

Alekh Agarwal,Yuda Song,Wen Sun,Kaiwen Wang,Mengdi Wang,Xuezhou Zhang

We study the problem of representational transfer in RL, where an agent first pretrains in a number of source tasks to discover a shared representation, which is subsequently used to learn a good policy in a target task. We propose a new notion of task relatedness between source and target tasks, and develop a novel approach for representational transfer under this assumption. Concretely, we show that given a generative access to source tasks, we can discover a representation, using which subsequent linear RL techniques quickly converge to a near-optimal policy, with only online access to the target task. The sample complexity is close to knowing the ground truth features in the target task, and comparable to prior representation learning results in the source tasks. We complement our positive results with lower bounds without generative access, and validate our findings with empirical evaluation on rich observation MDPs that require deep exploration.

****

Set Discrimination Contrastive Learning

Duong Hoang Le,Binh-Son Hua

In this work, we propose a self-supervised contrastive learning method that integrates the concept of set-based feature learning. The main idea of our method is to randomly construct sets of instances in a mini-batch and then learn to contrast the set representations. Inspired by set-based feature learning, we aggregate set features from individual sample features by a symmetric function. To improve the effectiveness of our set-based contrastive learning, we propose a set construction scheme built upon sample permutation in a mini-batch that allows a sample to appear in multiple sets, which naturally ensures common features among sets by construction. Our set construction scheme also increases both the number of positive and negative sets in a mini-batch, leading to better representation learning. We demonstrate the robustness of our method by seamlessly integrating it into existing contrastive learning methods such as SimCLR and MoCo. Extensive experiments demonstrate that our method consistently improves the performance of these contrastive learning methods in various datasets and downstream tasks.

****

What shapes the loss landscape of self supervised learning?

Liu Ziyin,Ekdeep Singh Lubana,Masahito Ueda,Hidenori Tanaka

Prevention of complete and dimensional collapse of representations has recently become a design principle for self-supervised learning (SSL). However, questions remain in our theoretical understanding: When do those collapses occur? What are the mechanisms and causes? We answer these questions by deriving and thoroughly analyzing an analytically tractable theory of SSL loss landscapes. In this theory, we identify the causes of the dimensional collapse and study the effect of normalization and bias. Finally, we leverage the interpretability afforded by the analytical theory to understand how dimensional collapse can be beneficial and what affects the robustness of SSL against data imbalance.

****

Learning Lightweight Object Detectors via Progressive Knowledge Distillation

Shengcao Cao,Mengtian Li,James Hays,Deva Ramanan,Yu-Xiong Wang,Liangyan Gui

Resource-constrained perception systems such as edge computing and vision-for-robotics require vision models to be both accurate and lightweight in computation and memory usage. Knowledge distillation is one effective strategy to improve the performance of lightweight classification models, but it is less well-explored for structured outputs such as object detection and instance segmentation, where the variable number of outputs and complex internal network modules complicate the distillation. In this paper, we propose a simple yet surprisingly effective sequential approach to knowledge distillation that progressively transfers the knowledge of a set of teachers to a given lightweight student. Our approach is inspired by curriculum learning: To distill knowledge from a highly accurate but complex teacher model, we construct a sequence of teachers to help the student gradually adapt. Our progressive distillation strategy can be easily combined with existing distillation mechanisms to consistently maximize student performance in various settings. To the best of our knowledge, we are the first to successfully distill knowledge from Transformer-based teacher detectors to convolution-based students, and unprecedentedly boost the performance of ResNet-50 based RetinaNet from 36.5% to 42.0% AP and Mask R-CNN from 38.2% to 42.5% AP on the MS COCO benchmark.

****

Topologically faithful image segmentation via induced matching of persistence barcodes

Nico Stucki,Johannes C. Paetzold,Suprosanna Shit,bjoern menze,Ulrich Bauer

Image segmentation is a largely researched field where neural networks find vast applications in many facets of technology.

Some of the most popular approaches to train segmentation networks employ loss functions optimizing pixel-overlap, an objective that is insufficient for many segmentation tasks. In recent years, their limitations fueled a growing interest in topology-aware methods, which aim to recover the correct topology of the segmented structures. However, so far, none of the existing approaches achieve a spatially correct matching between the topological features (persistence barcodes) of label (ground truth) and prediction (output of the neural network).

In this work, we propose the first topologically and feature-wise accurate metric and loss function for supervised image segmentation, which we term TopoMatch. We show how induced matchings guarantee the spatially correct matching between barcodes in a segmentation setting. Furthermore, we propose an efficient algorithm to compute TopoMatch for images. We show that TopoMatch is an interpretable metric to evaluate the topological correctness of segmentations. Moreover, we demonstrate how induced matchings can be used to train segmentation networks and improve the topological correctness of the segmentations across all 6 baseline datasets while preserving volumetric segmentation performance.

****

No Reason for No Supervision: Improved Generalization in Supervised Models

Mert Bülent Sar■y■ld■z,Yannis Kalantidis,Karteek Alahari,Diane Larlus

We consider the problem of training a deep neural network on a given classification task, e.g., ImageNet-1K (IN1K), so that it excels at both the training task as well as at other (future) transfer tasks. These two seemingly contradictory properties impose a trade-off between improving the model’s generalization and maintaining its performance on the original task. Models trained with self-supervised learning tend to generalize better than their supervised counterparts for transfer learning; yet, they still lag behind supervised models on IN1K. In this paper, we propose a supervised learning setup that leverages the best of both worlds. We extensively analyze supervised training using multi-scale crops for data augmentation and an expendable projector head, and reveal that the design of the projector allows us to control the trade-off between performance on the training task and transferability. We further replace the last layer of class weights with class prototypes computed on the fly using a memory bank and derive two models: t-ReX that achieves a new state of the art for transfer learning and outperforms top methods such as DINO and PAWS on IN1K, and t-ReX* that matches the highly optimized RSB-A1 model on IN1K while performing better on transfer tasks.
Code and pretrained models: https://europe.naverlabs.com/t-rex

****

Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies
Rui Yuan,Simon Shaolei Du,Robert M. Gower,Alessandro Lazaric,Lin Xiao
We consider infinite-horizon discounted Markov decision processes and study the convergence rates of the natural policy gradient (NPG) and the Q-NPG methods with the log-linear policy class. Using the compatible function approximation framework, both methods with log-linear policies can be written as approximate versions of the policy mirror descent (PMD) method. We show that both methods attain linear convergence rates and $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexities using a simple, non-adaptive geometrically increasing step size, without resorting to entropy or other strongly convex regularization. Lastly, as a byproduct, we obtain sublinear convergence rates for both methods with arbitrary constant step size.

****

Active Learning with Controllable Augmentation Induced Acquisition
Jianan Yang,Haobo Wang,Sai Wu,Gang Chen,Junbo Zhao
The mission of active learning is to iteratively identify the most informative data samples to annotate, and therefore to attain decent performance with much fewer samples. Despite the promise, the acquisition of informative unlabeled samples can be unreliable --- particularly during early cycles --- owning to limited data samples and sparse supervision. To tackle this, the data augmentation techniques seem straightforward yet promising to easily extend the exploration of the input space. In this work, we thoroughly study the coupling of data augmentation and active learning whereby we propose Controllable Augmentation ManiPulator for Active Learning. In contrast to the few prior work that touched on this line, CAMPAL emphasizes a tighten and better-controlled integration of data augmentation into the active learning framework, as in three folds: (i)-carefully designed data augmentation policies applied separately on labeled and unlabeled data pool in every cycle; (ii)-controlled and quantifiably optimizable augmentation strengths; (iii)-full but flexible coverage for most (if not all) active learning schemes. Through extensive empirical experiments, we bring the performance of active learning methods to a new level: an absolute performance boost of 16.99% on CIFAR-10 and 12.25% on SVHN with 1,000 annotated samples. Complementary to the empirical results, we further provide theoretical analysis and justification of CAMPAL.

****

Learning Axis-Aligned Decision Trees with Gradient Descent
Sascha Marton,Christian Bartelt,Stefan Lüdtke
Decision Trees are commonly used for many machine learning tasks due to their high interpretability. However, learning a decision tree from data is a difficult optimization problem, since it is non-convex and non-differentiable. Therefore, common approaches learn decision trees using a greedy growth algorithm that minimizes the impurity at each internal node. Unfortunately, this greedy procedure c

an lead to suboptimal trees.

In this paper, we present a novel approach for learning univariate, axis-aligned decision trees with gradient descent. This is achieved by applying backpropagation with an adjusted gradient flow on a dense decision tree representation that optimizes all decision tree parameters jointly. We show that our gradient-based optimization outperforms existing baselines on several binary classification benchmarks and achieves competitive results for multi-class tasks. To the best of our knowledge, this is the first approach that attempts to learn univariate, axis-aligned decision trees with gradient descent.

****

DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models
Cheng Lu,Yuhao Zhou,Fan Bao,Jianfei Chen,Chongxuan Li,Jun Zhu
Diffusion probabilistic models (DPMs) have achieved impressive success in high-resolution image synthesis, especially in recent large-scale text-to-image generation applications. An essential technique for improving the sample quality of DPMs is guided sampling, which usually needs a large guidance scale to obtain the best sample quality. The commonly-used fast sampler for guided sampling is DDIM, a first-order diffusion ODE solver that generally needs 100 to 250 steps for high-quality samples. Although recent works propose dedicated high-order solvers and achieve a further speedup for sampling without guidance, their effectiveness for guided sampling has not been well-tested before. In this work, we demonstrate that previous high-order fast samplers suffer from instability issues, and they even become slower than DDIM when the guidance scale grows large. To further speed up guided sampling, we propose DPM-Solver++, a high-order solver for the guided sampling of DPMs. DPM-Solver++ solves the diffusion ODE with the data prediction model and adopts thresholding methods to keep the solution matches training data distribution. We further propose a multistep variant of DPM-Solver++ to address the instability issue by reducing the effective step size. Experiments show that DPM-Solver++ can generate high-quality samples within only 15 to 20 steps for guided sampling by pixel-space and latent-space DPMs.

****

EVA3D: Compositional 3D Human Generation from 2D Image Collections
Fangzhou Hong,Zhaoxi Chen,Yushi LAN,Liang Pan,Ziwei Liu
Inverse graphics aims to recover 3D models from 2D observations. Utilizing differentiable rendering, recent 3D-aware generative models have shown impressive results of rigid object generation using 2D images. However, it remains challenging to generate articulated objects, like human bodies, due to their complexity and diversity in poses and appearances. In this work, we propose, EVA3D, an unconditional 3D human generative model learned from 2D image collections only. EVA3D can sample 3D humans with detailed geometry and render high-quality images (up to 512x256) without bells and whistles (e.g. super resolution). At the core of EVA3D is a compositional human NeRF representation, which divides the human body into local parts. Each part is represented by an individual volume. This compositional representation enables 1) inherent human priors, 2) adaptive allocation of network parameters, 3) efficient training and rendering. Moreover, to accommodate for the characteristics of sparse 2D human image collections (e.g. imbalanced pose distribution), we propose a pose-guided sampling strategy for better GAN learning. Extensive experiments validate that EVA3D achieves state-of-the-art 3D human generation performance regarding both geometry and texture quality. Notably, EVA3D demonstrates great potential and scalability to "inverse-graphics" diverse human bodies with a clean framework.

****

Spurious Local Minima Provably Exist for Deep Convolutional Neural Networks
Bo Liu,Keyi Fu,Tongtong Yuan
In this paper, we prove that a general family of spurious local minima exist in the loss landscape of deep convolutional neural networks with squared loss or cross-entropy loss. For this purpose, we develop some new techniques to solve the challenges introduced by convolutional layers. We solve a combinatorial problem

which considers the limited receptive fields of hidden neurons, and possible distinct activation status for different samples and different locations in feature maps, to show that a differentiation of data samples is always possible somewhere in feature maps. Training loss is then decreased by perturbation of network parameters that can affect different samples in different ways. Despite filters and biases are tied in each feature map, we give a construction in which this perturbation only affects the output of a single ReLU neuron and keeps the outputs at other locations unchanged. Finally, we give an example of nontrivial spurious local minimum in which different activation patterns of samples are explicitly constructed. Experimental results verify our theoretical findings.

****

Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game

Wei Xiong,Han Zhong,Chengshuai Shi,Cong Shen,Liwei Wang,Tong Zhang

Offline reinforcement learning (RL) aims at learning an optimal strategy using a pre-collected dataset without further interactions with the environment. While various algorithms have been proposed for offline RL in the previous literature, the minimax optimality has only been (nearly) established for tabular Markov decision processes (MDPs). In this paper, we focus on offline RL with linear function approximation and propose a new pessimism-based algorithm for offline linear MDP. At the core of our algorithm is the uncertainty decomposition via a reference function, which is new in the literature of offline RL under linear function approximation. Theoretical analysis demonstrates that our algorithm can match the performance lower bound up to logarithmic factors. We also extend our techniques to the two-player zero-sum Markov games (MGs), and establish a new performance lower bound for MGs, which tightens the existing result, and verifies the nearly minimax optimality of the proposed algorithm. To the best of our knowledge, these are the first computationally efficient and nearly minimax optimal algorithms for offline single-agent MDPs and MGs with linear function approximation.

****

Clustering Structure Identification With Ordering Graph

Zheng Xing,Weibing Zhao

In machine learning, data is often presented in the form of a graph or similarity (or distance) values between samples. Graph-based clustering methods such as spectral clustering are defined for general weighted graphs to identify the clustering structure. Graph construction research has developed significantly for decades, but the graph-based partition study still requires more attention because of its poor performance. For example, spectral clustering needs a post-processing (e.g., K-Means) step to uncover the clustering indicators. Yet, K-Means is sensitive to the initial center setting and easily falls into a local optimum. In this paper, we investigate a new type of graph-based clustering approach. Firstly, we introduce a new algorithm for the purpose of cluster analysis which does not explicitly produce a clustering of a dataset but instead creates an augmented graph representing its density-based ordered clustering structure. This ordered graph contains information equivalent to density-based clustering corresponding to a broad range of parameter settings. Secondly, we found that the graph matrix is shown in a block-diagonal form because of the nature of ordering. We propose a partition method to learn the graph matrix's block-diagonal structure and identify the clustering directly. The global optimality is guaranteed theoretically. We test the proposed method on synthetic datasets and five high-dimensional datasets. Experimental results show that the proposed method outperforms state-of-the-art graph-based clustering methods and improves their performance by roughly 10%-50%.

****

DaxBench: Benchmarking Deformable Object Manipulation with Differentiable Physics

Siwei Chen,Yiqing Xu,Cunjun Yu,Linfeng Li,Xiao Ma,Zhongwen Xu,David Hsu

Deformable object manipulation (DOM) is a long-standing challenge in robotics and has attracted significant interest recently. This paper presents DaXBench, a differentiable simulation framework for DOM. While existing work often focuses on a specific type of deformable objects, DaXBench supports fluid, rope, cloth ...; it provides a general-purpose benchmark to evaluate widely different DOM methods, including planning, imitation learning, and reinforcement learning. DaXBench combines recent advances in deformable object simulation with JAX, a high-performance computational framework. All DOM tasks in DaXBench are wrapped with the OpenAI Gym API for easy integration with DOM algorithms. We hope that DaXBench provides to the research community a comprehensive, standardized benchmark and a valuable tool to support the development and evaluation of new DOM methods. The code and video are available online.

****

Voxurf: Voxel-based Efficient and Accurate Neural Surface Reconstruction

Tong Wu,Jiaqi Wang,Xingang Pan,Xudong XU,Christian Theobalt,Ziwei Liu,Dahua Lin

Neural surface reconstruction aims to reconstruct accurate 3D surfaces based on multi-view images. Previous methods based on neural volume rendering mostly train a fully implicit model with MLPs, which typically require hours of training for a single scene. Recent efforts explore the explicit volumetric representation to accelerate the optimization via memorizing significant information with learnable voxel grids. However, existing voxel-based methods often struggle in reconstructing fine-grained geometry, even when combined with an SDF-based volume rendering scheme. We reveal that this is because 1) the voxel grids tend to break the color-geometry dependency that facilitates fine-geometry learning, and 2) the under-constrained voxel grids lack spatial coherence and are vulnerable to local minima. In this work, we present Voxurf, a voxel-based surface reconstruction approach that is both efficient and accurate. Voxurf addresses the aforementioned issues via several key designs, including 1) a two-stage training procedure that attains a coherent coarse shape and recovers fine details successively, 2) a dual color network that maintains color-geometry dependency, and 3) a hierarchical geometry feature to encourage information propagation across voxels. Extensive experiments show that Voxurf achieves high efficiency and high quality at the same time. On the DTU benchmark, Voxurf achieves higher reconstruction quality with a 20x training speedup compared to previous fully implicit methods. Our code is publicly available at https://github.com/wutong16/Voxurf/.

****

Conditional Positional Encodings for Vision Transformers

Xiangxiang Chu,Zhi Tian,Bo Zhang,Xinlong Wang,Chunhua Shen

We propose a conditional positional encoding (CPE) scheme for vision Transformers. Unlike previous fixed or learnable positional encodings that are predefined and independent of input tokens, CPE is dynamically generated and conditioned on the local neighborhood of the input tokens. As a result, CPE can easily generalize to the input sequences that are longer than what the model has ever seen during the training. Besides, CPE can keep the desired translation equivalence in vision tasks, resulting in improved performance. We implement CPE with a simple Position Encoding Generator (PEG) to get seamlessly incorporated into the current Transformer framework. Built on PEG, we present Conditional Position encoding Vision Transformer (CPVT). We demonstrate that CPVT has visually similar attention maps compared to those with learned positional encodings and delivers outperforming results.

****

Variational Reparametrized Policy Learning with Differentiable Physics

Zhiao Huang,Litian Liang,Zhan Ling,Xuanlin Li,Chuang Gan,Hao Su

We study the problem of policy parameterization for reinforcement learning (RL) with high-dimensional continuous action space. Our goal is to find a good way to parameterize the policy of continuous RL as a multi-modality distribution. To this end, we propose to treat the continuous RL policy as a generative model over the distribution of optimal trajectories. We use a diffusion process-like strategy to model the policy and derive a novel variational bound which is the optimization objective to learn the policy. To maximize the objective by gradient descent, we introduce the Reparameterized Policy Gradient Theorem. This theorem elegantly connects classical method REINFORCE and trajectory return optimization for computing the gradient of a policy. Moreover, our method enjoys strong explorat ion ability due to the multi-modality policy parameterization; notably, when a strong differentiable world model presents, our method also enjoys the fast convergence speed of trajectory optimization. We evaluate our method on numerical problems and manipulation tasks within a differentiable simulator. Qualitative results show its ability to capture the multi-modality distribution of optimal trajectories, and quantitative results show that it can avoid local optima and outperforms baseline approaches.

****

GENERALIZED MATRIX LOCAL LOW RANK REPRESENTATION BY RANDOM PROJECTION AND SUBMATRIX PROPAGATION

Pengtao Dang,Wennan Chang,Haiqi Zhu,Changlin Wan,Tong Zhao,Tingo Guo,Paul Salama,Sha Cao,Chi Zhang

Detecting distinct submatrices of low rank property is a highly desirable matrix representation learning technique for the ease of data interpretation, called the matrix local low rank representation (MLLRR). Based on different mathematical assumptions of the local pattern, the MLLRR problem could be categorized into two sub-problems, namely local constant variation (LCV) and local linear low rank (LLR). Existing solutions on MLLRR only focused on the LCV problem, which misses a substantial amount of true and interesting patterns. In this work, we develop a novel matrix computational framework called RPSP (Random Probing based submatrix Propagation) that provides an effective solution for both of the LCV and LLR problems. RPSP detects local low rank patterns that grow from small submatrices of low rank property, which are determined by a random projection approach. RPSP is supported by theories of random projection. Experiments on synthetic data demonstrate that RPSP outperforms all state-of-the-art methods, with the capacity to robustly and correctly identify the low rank matrices under both LCV and LLR settings. On real-world datasets, RPSP also demonstrates its effectiveness in identifying interpretable local low rank matrices.

****

Stable, Efficient, and Flexible Monotone Operator Implicit Graph Neural Networks

Justin Baker,Qingsong Wang,Bao Wang

Implicit graph neural networks (IGNNs) that solve a fixed-point equilibrium equation for representation learning can learn the long-range dependencies (LRD) in the underlying graphs and show remarkable performance for various graph learning tasks. However, the expressivity of IGNNs is limited by the constraints for their well-posedness guarantee. Moreover, when IGNNs become effective for learning LRD, their eigenvalues converge to the value that slows down the convergence, and their performance is unstable across different tasks. In this paper, we provide a new well-posedness condition of IGNNs leveraging monotone operator theory. The new well-posedness characterization informs us to design effective parameterizations to improve the accuracy, efficiency, and stability of IGNNs. Leveraging accelerated operator splitting schemes and graph diffusion convolution, we design efficient and flexible implementations of monotone operator IGNNs that are significantly faster and more accurate than existing IGNNs.

****

ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills

Jiayuan Gu,Fanbo Xiang,Xuanlin Li,Zhan Ling,Xiqiang Liu,Tongzhou Mu,Yihe Tang,Stone Tao,Xinyue Wei,Yunchao Yao,Xiaodi Yuan,Pengwei Xie,Zhiao Huang,Rui Chen,Hao Su

Generalizable manipulation skills, which can be composed to tackle long-horizon and complex daily chores, are one of the cornerstones of Embodied AI. However, existing benchmarks, mostly composed of a suite of simulatable environments, are insufficient to push cutting-edge research works because they lack object-level topological and geometric variations, are not based on fully dynamic simulation, or are short of native support for multiple types of manipulation tasks. To this end, we present ManiSkill2, the next generation of the SAPIEN ManiSkill benchmark, to address critical pain points often encountered by researchers when using benchmarks for generalizable manipulation skills. ManiSkill2 includes 20 manipulation task families with 2000+ object models and 4M+ demonstration frames, whic

h cover stationary/mobile-base, single/dual-arm, and rigid/soft-body manipulation tasks with 2D/3D-input data simulated by fully dynamic engines. It defines a unified interface and evaluation protocol to support a wide range of algorithms (e.g., classic sense-plan-act, RL, IL), visual observations (point cloud, RGBD), and controllers (e.g., action type and parameterization). Moreover, it empowers fast visual input learning algorithms so that a CNN-based policy can collect samples at about 2000 FPS with 1 GPU and 16 processes on a regular workstation. It implements a render server infrastructure to allow sharing rendering resources across all environments, thereby significantly reducing memory usage. We open-source all codes of our benchmark (simulator, environments, and baselines) and host an online challenge open to interdisciplinary researchers.
****

Deja Vu: Continual Model Generalization for Unseen Domains
Chenxi Liu,Lixu Wang,Lingjuan Lyu,Chen Sun,Xiao Wang,Qi Zhu
In real-world applications, deep learning models often run in non-stationary environments where the target data distribution continually shifts over time. There have been numerous domain adaptation (DA) methods in both online and offline modes to improve cross-domain adaptation ability. However, these DA methods typically only provide good performance after a long period of adaptation, and perform poorly on new domains before and during adaptation – in what we call the “Unfamiliar Period”, especially when domain shifts happen suddenly and significantly. On the other hand, domain generalization (DG) methods have been proposed to improve the model generalization ability on unadapted domains. However, existing DG works are ineffective for continually changing domains due to severe catastrophic forgetting of learned knowledge. To overcome these limitations of DA and DG in handling the Unfamiliar Period during continual domain shift, we propose RaTP, a framework that focuses on improving models’ target domain generalization (TDG) capability, while also achieving effective target domain adaptation (TDA) capability right after training on certain domains and forgetting alleviation (FA) capability on past domains. RaTP includes a training-free data augmentation module to prepare data for TDG, a novel pseudo-labeling mechanism to provide reliable supervision for TDA, and a prototype contrastive alignment algorithm to align different domains for achieving TDG, TDA and FA. Extensive experiments on Digits, PACS, and DomainNet demonstrate that RaTP significantly outperforms state-of-the-art works from Continual DA, Source-Free DA, Test-Time/Online DA, Single DG, Multiple DG and Unified DA&DG in TDG, and achieves comparable TDA and FA capabilities.
****

A Graph Neural Network Approach to Automated Model Building in Cryo-EM Maps
Kiarash Jamali,Dari Kimanius,Sjors HW Scheres
Electron cryo-microscopy (cryo-EM) produces three-dimensional (3D) maps of the electrostatic potential of biological macromolecules, including proteins. At sufficient resolution, the cryo-EM maps, along with some knowledge about the imaged molecules, allow de novo atomic modelling. Typically, this is done through a laborious manual process. Recent advances in machine learning applications to protein structure prediction show potential for automating this process. Taking inspiration from these techniques, we have built ModelAngelo for automated model building of proteins in cryo-EM maps. ModelAngelo first uses a residual convolutional neural network (CNN) to initialize a graph representation with nodes assigned to individual amino acids of the proteins in the map and edges representing the protein chain. The graph is then refined with a graph neural network (GNN) that combines the cryo-EM data, the amino acid sequence data and prior knowledge about protein geometries. The GNN refines the geometry of the protein chain and classifies the amino acids for each of its nodes. The final graph is post-processed with a hidden Markov model (HMM) search to map each protein chain to entries in a user provided sequence file. Application to 28 test cases shows that ModelAngelo outperforms state-of-the-art and approximates manual building for cryo-EM maps with resolutions better than 3.5 A.
****

Stealing and Defending Transformer-based Encoders

Adam Dziedzic,Franziska Boenisch,Mingjian Jiang,Haonan Duan,Nicolas Papernot

Self-supervised learning (SSL) has become the predominant approach to training on large amounts of unlabeled data. New real-world APIs offer services to generate high-dimensional representations for given inputs based on SSL encoders with transformer architectures. Recent efforts highlight that it is possible to steal high-quality SSL encoders trained on convolutional neural networks. In this work, we are the first to extend this line of work to stealing and defending transformer-based encoders in both language and vision domains. We show that it is possible to steal transformer-based sentence embedding models solely using their returned representations and with 40x fewer queries than the number of victim's training data points. We also decrease the number of required stealing queries for the vision encoders by leveraging semi-supervised learning. Finally, to defend vision transformers against stealing attacks, we propose a defense technique that combines watermarking with dataset inference. Our method creates a unique encoder signature based on a private data subset that acts as a secret seed during training. By applying dataset inference on the seed, we can then successfully identify stolen transformers.

****

On the Lower Bound of Minimizing Polyak-■ojasiewicz functions

Pengyun Yue,Cong Fang,Zhouchen Lin

Polyak-■ojasiewicz (PL) [Polyak, 1963] condition is a weaker condition than the strong convexity but suffices to ensure a global convergence for the Gradient Descent algorithm. In this paper, we study the lower bound of algorithms using first-order oracles to find an approximate optimal solution. We show that any first-order algorithm requires at least $\Omega\left((L/\mu)^{1-\alpha} \right)$ gradient costs to find an ϵ-approximate optimal solution for a general L-smooth function that has an μ-PL constant for any $\alpha>0$. This result demonstrates the near optimality of the Gradient Descent algorithm to minimize smooth PL functions in the sense that there exists a ``hard'' PL function such that no first-order algorithm can be faster by a polynomial order. In contrast, it is well-known that the momentum technique, e.g. [Nesterov, 2003, chap. 2] can provably accelerate Gradient Descent to $O\left(\sqrt{L/\hat{\mu}}\log\frac{1}{\epsilon}\right)$ gradient costs for functions that are L-smooth and $\hat{\mu}$-strongly convex. Therefore, our result distinguishes the hardness of minimizing a smooth PL function and a smooth strongly convex function as the complexity of the former cannot be improved by any polynomial order in general.

****

VectorMapNet: End-to-end Vectorized HD Map Learning

Yicheng Liu,Tianyuan Yuan,Yue Wang,Yilun Wang,Hang Zhao

Autonomous driving systems require a good understanding of surrounding environments, including moving obstacles and static High-Definition (HD) semantic map elements. Existing methods approach the semantic map problem by offline manual annotation, which suffers from serious scalability issues. Recent learning-based methods produce dense rasterized segmentation predictions to construct maps. However, these predictions do not include instance information of individual map elements and require heuristic post-processing to obtain vectorized maps. To tackle these challenges, we introduce an end-to-end vectorized HD map learning pipeline, termed VectorMapNet. VectorMapNet takes onboard sensor observations and predicts a sparse set of polylines in the bird's-eye view. This pipeline can explicitly model the spatial relation between map elements and generate vectorized maps that are friendly to downstream autonomous driving tasks. Extensive experiments show that VectorMapNet achieve strong map learning performance on both nuScenes and Argoverse2 dataset, surpassing previous state-of-the-art methods by 14.2 mAP and 14.6mAP. Qualitatively, we also show that VectorMapNet is capable of generating comprehensive maps and capturing more fine-grained details of road geometry.

To the best of our knowledge, VectorMapNet is the first work designed towards end-to-end vectorized map learning from onboard sensors.

****

An information-theoretic approach to unsupervised keypoint representation learni

ng
Ali Younes,Simone Schaub-Meyer,Georgia Chalvatzaki

Extracting informative representations from videos is fundamental for the effective learning of various downstream tasks. Inspired by classical works on saliency, we present a novel information-theoretic approach to discover meaningful representations from videos in an unsupervised fashion. We argue that local entropy of pixel neighborhoods and its evolution in a video stream is a valuable intrinsic supervisory signal for learning to attend to salient features. We, thus, abstract visual features into a concise representation of keypoints that serve as dynamic information transporters. We discover in an unsupervised fashion spatio-temporally consistent keypoint representations that carry the prominent information across video frames, thanks to two original information-theoretic losses. First, a loss that maximizes the information covered by the keypoints in a frame. Second, a loss that encourages optimized keypoint transportation over time, thus, imposing consistency of the information flow. We evaluate our keypoint-based representation compared to state-of-the-art baselines in different downstream tasks such as learning object dynamics. To evaluate the expressivity and consistency of the keypoints, we propose a new set of metrics. Our empirical results showcase the superior performance of our information-driven keypoints that resolve challenges like attendance to both static and dynamic objects, and to objects abruptly entering and leaving the scene.

****

Distilling Cognitive Backdoor Patterns within an Image
Hanxun Huang,Xingjun Ma,Sarah Monazam Erfani,James Bailey

This paper proposes a simple method to distill and detect backdoor patterns within an image: \emph{Cognitive Distillation} (CD). The idea is to extract the ``minimal essence" from an input image responsible for the model's prediction. CD optimizes an input mask to extract a small pattern from the input image that can lead to the same model output (i.e., logits or deep features). The extracted pattern can help understand the cognitive mechanism of a model on clean vs. backdoor images and is thus called a \emph{Cognitive Pattern} (CP). Using CD and the distilled CPs, we uncover an interesting phenomenon of backdoor attacks: despite the various forms and sizes of trigger patterns used by different attacks, the CPs of backdoor samples are all surprisingly and suspiciously small.

One thus can leverage the learned mask to detect and remove backdoor examples from poisoned training datasets.

We conduct extensive experiments to show that CD can robustly detect a wide range of advanced backdoor attacks.

We also show that CD can potentially be applied to help detect potential biases from face datasets.

Code is available at https://github.com/HanxunH/CognitiveDistillation.

****

Curriculum Reinforcement Learning via Morphology-Environment Co-Evolution
Shuang Ao,Tianyi Zhou,Guodong Long,Jing Jiang

Throughout long history, natural species have learned to survive by evolving their physical structures adaptive to the environment changes. In contrast, current reinforcement learning (RL) studies mainly focus on training an agent with a fixed morphology (e.g., skeletal structure and joint attributes) in a fixed environment, which can hardly generalize to changing environments or new tasks. In this paper, we optimize an RL agent and its morphology through ''morphology-environment co-evolution (MECE)'', in which the morphology keeps being updated to adapt to the changing environment, while the environment is modified progressively to bring new challenges and stimulate the improvement of the morphology. This leads to a curriculum to train generalizable RL, whose morphology and policy are optimized for different environments. Instead of hand-crafting the curriculum, we train two policies to automatically change the morphology and the environment. To this end, (1) we develop two novel and effective rewards for the two policies, which are solely based on the learning dynamics of the RL agent; (2) we design a scheduler to automatically determine when to change the environment and the morphology. We find these two designs are critical to the success of MECE, as veri fied by extensive ablation studies. In experiments on two classes of tasks, the morphology and RL policies trained via MECE exhibit significantly better generalization performance in unseen test environments than SOTA morphology optimization methods. Our ablation studies on the two MECE policies further show that the co-evolution between the morphology and environment is the key to the success.

****

Domain Generalization with Small Data

Kecheng Chen,Elena Gal,Hong Yan,Haoliang Li

In this work, we propose to tackle the problem of domain generalization in the context of insufficient samples. Instead of extracting latent feature embeddings based on deterministic models, we propose to learn a domain-invariant representation based on the probabilistic framework by mapping each data point into probabilistic embeddings. Specifically, we first extend empirical maximum mean discrepancy (MMD) to a novel probabilistic MMD that can measure the discrepancy between mixture distributions (i.e., source domains) consisted of a serial of latent distributions rather than latent points. Moreover, instead of imposing the contrastive semantic alignment (CSA) loss based on pairs of latent points, a novel probabilistic CSA loss encourages positive probabilistic embedding pairs to be closer while pulling other negative ones apart. Benefiting from the learned representation captured by probabilistic models, our proposed method can marriage the measurement on the distribution over distributions (i.e., the global perspective alignment) and the distribution-based contrastive semantic alignment (i.e., the local perspective alignment). Extensive experimental results on three challenging medical datasets show the effectiveness of our proposed method in the context of insufficient data compared with state-of-the-art baseline methods.

****

3D generation on ImageNet

Ivan Skorokhodov,Aliaksandr Siarohin,Yinghao Xu,Jian Ren,Hsin-Ying Lee,Peter Wonka,Sergey Tulyakov

All existing 3D-from-2D generators are designed for well-curated single-category datasets, where all the objects have (approximately) the same scale, 3D location, and orientation, and the camera always points to the center of the scene. This makes them inapplicable to diverse, in-the-wild datasets of non-alignable scenes rendered from arbitrary camera poses. In this work, we develop a 3D generator with Generic Priors (3DGP): a 3D synthesis framework with more general assumptions about the training data, and show that it scales to very challenging datasets, like ImageNet. Our model is based on three new ideas. First, we incorporate an inaccurate off-the-shelf depth estimator into 3D GAN training via a special depth adaptation module to handle the imprecision. Then, we create a flexible camera model and a regularization strategy for it to learn its distribution parameters during training. Finally, we extend the recent ideas of transferring knowledge from pretrained classifiers into GANs for patch-wise trained models by employing a simple distillation-based technique on top of the discriminator. It achieves more stable training than the existing methods and speeds up the convergence by at least 40%. We explore our model on four datasets: SDIP Dogs 256^2, SDIP Elephants 256^2, LSUN Horses 256^2, and ImageNet 256^2 and demonstrate that 3DGP outperforms the recent state-of-the-art in terms of both texture and geometry quality. Code and visualizations: https://snap-research.github.io/3dgp.

****

Revocable Deep Reinforcement Learning with Affinity Regularization for Outlier-Robust Graph Matching

Chang Liu,Zetian Jiang,Runzhong Wang,Lingxiao Huang,Pinyan Lu,Junchi Yan

Graph matching (GM) has been a building block in various areas including computer vision and pattern recognition. Despite recent impressive progress, existing deep GM methods often have obvious difficulty in handling outliers, which are ubiquitous in practice. We propose a deep reinforcement learning based approach RGM, whose sequential node matching scheme naturally fits the strategy for selective inlier matching against outliers. A revocable action framework is devised to improve the agent's flexibility against the complex constrained GM. Moreover, we propose a quadratic approximation technique to regularize the affinity score, in the presence of outliers. As such, the agent can finish inlier matching timely when the affinity score stops growing, for which otherwise an additional parameter i.e. the number of inliers is needed to avoid matching outliers. In this paper, we focus on learning the back-end solver under the most general form of GM: the Lawler's QAP, whose input is the affinity matrix. Especially, our approach can also boost existing GM methods that use such input. Experiments on multiple real-world datasets demonstrate its performance regarding both accuracy and robustness.

****

Hierarchical Prompting Improves Visual Recognition On Accuracy, Data Efficiency and Explainability

Wenhao Wang,Yifan Sun,Wei Li,Yi Yang

When humans try to distinguish some inherently similar visual concepts, e.g., Rosa Peace and China Rose, they may use the underlying hierarchical taxonomy to prompt the recognition. For example, given a prompt that the image belongs to the rose family, a person can narrow down the category range and thus focuses on the comparison between different roses. In this paper, we explore the hierarchical prompting for deep visual recognition (image classification, in particular) based on the prompting mechanism of the transformer. We show that the transformer can take the similar benefit by injecting the coarse-class prompts into the intermediate blocks. The resulting Transformer with Hierarchical Prompting (TransHP) is very simple and consists of three steps: 1) TransHP learns a set of prompt tokens to represent the coarse classes, 2) learns to predict the coarse class of the input image using an intermediate block, and 3) absorbs the prompt token of the predicted coarse class into the feature tokens. Consequently, the injected coarse-class prompt conditions (influences) the subsequent feature extraction and encourages better focus on the relatively subtle differences among the descendant classes. Through extensive experiments on popular image classification datasets, we show that this simple hierarchical prompting improves visual recognition on classification accuracy (e.g., improving ViT-B/16 by $+2.83\%$ ImageNet classification accuracy), training data efficiency (e.g., $+12.69\%$ improvement over the baseline under 10% ImageNet training data), and model explainability.

****

Convergence of the mini-batch SIHT algorithm

Saeed Damadi,Jinglai Shen

The Iterative Hard Thresholding (IHT) algorithm has been considered extensively as an effective deterministic algorithm for solving sparse optimizations.

The IHT algorithm benefits from the information of the batch (full) gradient at each point and this information is a crucial key for the convergence analysis of the generated sequence. However, this strength becomes a weakness when it comes to machine learning and high dimensional statistical applications because calculating the batch gradient at each iteration is computationally expensive or impractical. Fortunately, in these applications the objective function has a summation structure that can be taken advantage of to approximate the batch gradient by the stochastic mini-batch gradient. In this paper, we study the mini-batch Stochastic IHT (SIHT) algorithm for solving the sparse optimizations. As opposed to previous works where increasing and variable mini-batch size is necessary for derivation, we fix the mini-batch size according to a lower bound that we derive and show our work. To prove stochastic convergence of the objective value function we first establish a critical sparse stochastic gradient descent property. Using this stochastic gradient descent property we show that the sequence generated by the stochastic mini-batch SIHT is a supermartingale sequence and converges with probability one. Unlike previous work we do not assume the function to be a restricted strongly convex. To the best of our knowledge, in the regime of sparse optimization, this is the first time in the literature that it is shown that the sequence of the stochastic function values converges with probability one by fixing the mini-batch size for all steps.

****

Decomposing Texture and Semantics for Out-of-distribution Detection

Jeong-Hyeon Moon,Namhyuk Ahn,Kyung-Ah Sohn

Out-of-distribution (OOD) detection has made significant progress in recent years because the distribution mismatch between the training and testing can severely deteriorate the reliability of a machine learning system.Nevertheless, the lack of precise interpretation of the in-distribution limits the application of OOD detection methods to real-world system pipielines. To tackle this issue, we decompose the definition of the in-distribution into texture and semantics, motivated by real-world scenarios. In addition, we design new benchmarks to measure the robustness that OOD detection methods should have. To achieve a good balance between the OOD detection performance and robustness, our method takes a divide-and-conquer approach. That is, the model first tackles each component of the texture and semantics separately, and then combines them later. Such design philosophy is empirically proven by a series of benchmarks including not only ours but also the conventional counterpart.
****
Rethinking the Expressive Power of GNNs via Graph Biconnectivity
Bohang Zhang,Shengjie Luo,Liwei Wang,Di He
Designing expressive Graph Neural Networks (GNNs) is a central topic in learning graph-structured data. While numerous approaches have been proposed to improve GNNs with respect to the Weisfeiler-Lehman (WL) test, for most of them, there is still a lack of deep understanding of what additional power they can systematically and provably gain. In this paper, we take a fundamentally different perspective to study the expressive power of GNNs beyond the WL test. Specifically, we introduce a novel class of expressivity metrics via graph biconnectivity and highlight their importance in both theory and practice. As biconnectivity can be easily calculated using simple algorithms that have linear computational costs, it is natural to expect that popular GNNs can learn it easily as well. However, after a thorough review of prior GNN architectures, we surprisingly find that most of them are not expressive for any of these metrics. The only exception is the ESAN framework (Bevilacqua et al., 2022), for which we give a theoretical justification of its power. We proceed to introduce a principled and more efficient approach, called the Generalized Distance Weisfeiler-Lehman (GD-WL), which is provably expressive for all biconnectivity metrics. Practically, we show GD-WL can be implemented by a Transformer-like architecture that preserves expressiveness and enjoys full parallelizability. A set of experiments on both synthetic and real datasets demonstrates that our approach can consistently outperform prior GNN architectures.
****
One Transformer Can Understand Both 2D & 3D Molecular Data
Shengjie Luo,Tianlang Chen,Yixian Xu,Shuxin Zheng,Tie-Yan Liu,Liwei Wang,Di He
Unlike vision and language data which usually has a unique format, molecules can naturally be characterized using different chemical formulations. One can view a molecule as a 2D graph or define it as a collection of atoms located in a 3D space. For molecular representation learning, most previous works designed neural networks only for a particular data format, making the learned models likely to fail for other data formats. We believe a general-purpose neural network model for chemistry should be able to handle molecular tasks across data modalities. To achieve this goal, in this work, we develop a novel Transformer-based Molecular model called Transformer-M, which can take molecular data of 2D or 3D formats as input and generate meaningful semantic representations. Using the standard Transformer as the backbone architecture, Transformer-M develops two separated channels to encode 2D and 3D structural information and incorporate them with the atom features in the network modules. When the input data is in a particular format, the corresponding channel will be activated, and the other will be disabled. By training on 2D and 3D molecular data with properly designed supervised signals, Transformer-M automatically learns to leverage knowledge from different data modalities and correctly capture the representations. We conducted extensive experiments for Transformer-M. All empirical results show that Transformer-M can simultaneously achieve strong performance on 2D and 3D tasks, suggesting its broad applicability. The code and models will be made publicly available at https://github.com/lsj2408/Transformer-M.

****

Generating Diverse Cooperative Agents by Learning Incompatible Policies

Rujikorn Charakorn,Poramate Manoonpong,Nat Dilokthanakul

Training a robust cooperative agent requires diverse partner agents. However, obtaining those agents is difficult. Previous works aim to learn diverse behaviors by changing the state-action distribution of agents. But, without information about the task's goal, the diversified agents are not guided to find other important, albeit sub-optimal, solutions: the agents might learn only variations of the same solution. In this work, we propose to learn diverse behaviors via policy compatibility. Conceptually, policy compatibility measures whether policies of interest can coordinate effectively. We theoretically show that incompatible policies are not similar. Thus, policy compatibility—which has been used exclusively as a measure of robustness—can be used as a proxy for learning diverse behaviors. Then, we incorporate the proposed objective into a population-based training scheme to allow concurrent training of multiple agents. Additionally, we use state-action information to induce local variations of each policy. Empirically, the proposed method consistently discovers more solutions than baseline methods across various multi-goal cooperative environments. Finally, in multi-recipe Overcooked, we show that our method produces populations of behaviorally diverse agents, which enables generalist agents trained with such a population to be more robust.

See our project page at https://bit.ly/marl-lipo

****

Mind the Gap: Offline Policy Optimization for Imperfect Rewards

Jianxiong Li,Xiao Hu,Haoran Xu,Jingjing Liu,Xianyuan Zhan,Qing-Shan Jia,Ya-Qin Zhang

Reward function is essential in reinforcement learning (RL), serving as the guiding signal to incentivize agents to solve given tasks, however, is also notoriously difficult to design. In many cases, only imperfect rewards are available, which inflicts substantial performance loss for RL agents. In this study, we propose a unified offline policy optimization approach, \textit{RGM (Reward Gap Minimization)}, which can smartly handle diverse types of imperfect rewards. RGM is formulated as a bi-level optimization problem: the upper layer optimizes a reward correction term that performs visitation distribution matching w.r.t. some expert data; the lower layer solves a pessimistic RL problem with the corrected rewards. By exploiting the duality of the lower layer, we derive a tractable algorithm that enables sampled-based learning without any online interactions. Comprehensive experiments demonstrate that RGM achieves superior performance to existing methods under diverse settings of imperfect rewards. Further, RGM can effectively correct wrong or inconsistent rewards against expert preference and retrieve useful information from biased rewards. Code is available at https://github.com/Facebear-ljx/RGM.

****

Gamma Sampling: Fine-grained Controlling Language Models without Training

Shangda Wu,Maosong Sun

The dominant approaches for controlling language models achieve prominence in controlling high-level attributes (e.g. topic and sentiment). However, these methods often require condition-specific data or are computationally expensive. We propose a new simple guided decoding method, Gamma Sampling, which does not require any training data to achieve fine-grained controllable text generation while maintaining a fast generation speed. Gamma Sampling introduces attribute-related information (provided by humans or language models themselves) into the sampling process to guide language models to generate texts with desired attributes. Since no training is involved, Gamma Sampling can be easily applied to any language model for controllable text generation. Through experiments, we show that Gamma Sampling-steered GPT2-small (117M) outperforms baselines such as PPLM (345M) and CTRL (1.6B) in diversity, attribute relevance, and overall quality of generated samples.

****

Label Distribution Learning via Implicit Distribution Representation

Zhuoran Zheng,Xiuyi Jia

In contrast to multi-label learning, label distribution learning characterizes the polysemy of examples by a label distribution to represent richer semantics. In the learning process of label distribution, the training data is collected mainly by manual annotation or label enhancement algorithms to generate label distribution. Unfortunately, the complexity of the manual annotation task or the inaccuracy of the label enhancement algorithm leads to noise and uncertainty in the label distribution training set. To alleviate this problem, we introduce the implicit distribution in the label distribution learning framework to characterize the uncertainty of each label value. Specifically, we use deep implicit representation learning to construct a label distribution matrix with Gaussian prior constraints, where each row component corresponds to the distribution estimate of each label value, and this row component is constrained by a prior Gaussian distribution to moderate the noise and uncertainty interference of the label distribution dataset. Finally, each row component of the label distribution matrix is transformed into a standard label distribution form by using the self-attention algorithm. In addition, some approaches with regularization characteristics are conducted in the training phase to improve the performance of the model.

****

MultiWave: Multiresolution Deep Architectures through Wavelet Decomposition for Multivariate Timeseries Forecasting and Prediction

Iman Deznabi,Madalina Fiterau

One of the challenges in multivariate time series modeling is that changes in signals occur with different frequencies, even when the sampling rate is consistent across signals. In the case of multivariate time series prediction, the outcome is also determined by patterns of different frequencies. These encapsulate both long-term and short-term effects, which have so far not been sufficiently leveraged by deep learning time series models. We fill this gap by introducing a framework, called MultiWave, which augments any deep learning time series model with components operating at the intrinsic frequencies of the signals. MultiWave applies wavelet decomposition on each signal to obtain subsignals of different frequencies and groups all subsignals in the same frequency band together to train a component. The output of the components is combined through a gating mechanism that removes irrelevant frequencies for the given predictive task. We show that MultiWave accurately determines the informative frequency bands and that the augmented models including components trained to operate on those bands outperform the original models. We further show that applying MultiWave on top of different deep learning models improves their performance in several real-world applications.

****

Learning to Compose Soft Prompts for Compositional Zero-Shot Learning

Nihal V. Nayak,Peilin Yu,Stephen Bach

We introduce compositional soft prompting (CSP), a parameter-efficient learning technique to improve the zero-shot compositionality of large-scale pretrained vision-language models (VLMs) like CLIP. We develop CSP for compositional zero-shot learning, the task of predicting unseen attribute-object compositions (e.g., old cat and young tiger). VLMs have a flexible text encoder that can represent arbitrary classes as natural language prompts but they often underperform task-specific architectures on the compositional zero-shot benchmark datasets. CSP treats the attributes and objects that define classes as learnable tokens of vocabulary. During training, the vocabulary is tuned to recognize classes that compose tokens in multiple ways (e.g., old cat and white cat). At test time, we recompose the learned attribute-object vocabulary in new combinations to recognize novel classes. We show that CSP outperforms the CLIP on benchmark datasets by an average of 10.9 percentage points on AUC. CSP also outperforms CoOp, a soft prompting method that fine-tunes the prefix context tokens, by an average of 5.8 percentage points on AUC. We perform additional experiments to show that CSP improves generalization to higher-order attribute-attribute-object compositions (e.g., old white cat) and combinations of pretrained attributes and fine-tuned objects. The code is available at https://github.com/BatsResearch/csp.

****

SQA3D: Situated Question Answering in 3D Scenes

Xiaojian Ma,Silong Yong,Zilong Zheng,Qing Li,Yitao Liang,Song-Chun Zhu,Siyuan Huang

We propose a new task to benchmark scene understanding of embodied agents: Situated Question Answering in 3D Scenes (SQA3D). Given a scene context (e.g., 3D scan), SQA3D requires the tested agent to first understand its situation (position, orientation, etc.) in the 3D scene as described by text, then reason about its surrounding environment and answer a question under that situation. Based upon 650 scenes from ScanNet, we provide a dataset centered around 6.8k unique situations, along with 20.4k descriptions and 33.4k diverse reasoning questions for these situations. These questions examine a wide spectrum of reasoning capabilities for an intelligent agent, ranging from spatial relation comprehension to commonsense understanding, navigation, and multi-hop reasoning. SQA3D imposes a significant challenge to current multi-modal especially 3D reasoning models. We evaluate various state-of-the-art approaches and find that the best one only achieves an overall score of 47.20%, while amateur human participants can reach 90.06%. We believe SQA3D could facilitate future embodied AI research with stronger situation understanding and reasoning capability.

****

The Benefits of Model-Based Generalization in Reinforcement Learning

Kenny John Young,Aditya Ramesh,Louis Kirsch,Jürgen Schmidhuber

Model-Based Reinforcement Learning (RL) is widely believed to have the potential to improve sample efficiency by allowing an agent to synthesize large amounts of imagined experience. Experience Replay (ER) can be considered a simple kind of model, which has proved extremely effective at improving the stability and efficiency of deep RL. In principle, a learned parametric model could improve on ER by generalizing from real experience to augment the dataset with additional plausible experience. However, owing to the many design choices involved in empirically successful algorithms, it can be very hard to establish where the benefits are actually coming from. Here, we provide theoretical and empirical insight into when, and how, we can expect data generated by a learned model to be useful. First, we provide a general theorem motivating how learning a model as an intermediate step can narrow down the set of possible value functions more than learning a value function directly from data using the Bellman equation. Second, we provide an illustrative example showing empirically how a similar effect occurs in a more concrete setting with neural network function approximation. Finally, we provide extensive experiments showing the benefit of model-based learning for online RL in environments with combinatorial complexity, but factored structure that allows a learned model to generalize. In these experiments, we take care to control for other factors in order to isolate, insofar as possible, the benefit of using experience generated by a learned model relative to ER alone.

****

Revisiting Higher-Order Gradient Methods for Multi-Agent Reinforcement Learning

Ariyan Bighashdel,Daan De Geus,Pavol Jancura,Gijs Dubbelman

This paper revisits Higher-Order Gradient (HOG) methods for Multi-Agent Reinforcement Learning (MARL). HOG methods are algorithms in which agents use higher-order gradient information to account for other agents' anticipated learning, and are shown to improve coordination in games with self-interested agents. So far, however, HOG methods are only applied to games with low-dimensional state spaces due to inefficient computation and preservation of higher-order gradient information. In this work, we solve these limitations and propose a HOG framework that can be applied to games with higher-dimensional state spaces. Moreover, we show that current HOG methods, when applied to games with common-interested agents, i.e., team games, can lead to miscoordination between the agents. To solve this, we propose Hierarchical Reasoning (HR) to improve coordination in team games, and we experimentally show that our proposed HR significantly outperforms state-of-the-art methods in standard multi-agent games. With our contributions, we great

ly improve the applicability of HOG methods for MARL. For reproducibility, the code used for our work will be shared after the reviewing process.
****
Efficient Covariance Estimation for Sparsified Functional Data
Sijie Zheng,Fandong Meng,Jie Zhou
To avoid prohibitive computation cost of sending entire data, we propose four sparsification schemes Random-knots, Random-knots-Spatial, B-spline, Bspline-Spatial, and present corresponding nonparametric estimation of the covariance function. The covariance estimators are asymptotically equivalent to the sample covariance computed directly from the original data. And the estimated functional principal components effectively approximate the infeasible principal components under regularity conditions. The convergence rate reflects that leveraging spatial correlation and B-spline interpolation helps to reduce information loss. Data-driven selection method is further applied to determine the number of eigenfunctions in the model. Extensive numerical experiments are conducted to illustrate the theoretical results.
****
Sparse Mixture-of-Experts are Domain Generalizable Learners
Bo Li,Yifei Shen,Jingkang Yang,Yezhen Wang,Jiawei Ren,Tong Che,Jun Zhang,Ziwei Liu
Human visual perception can easily generalize to out-of-distributed visual data, which is far beyond the capability of modern machine learning models. Domain generalization (DG) aims to close this gap, with existing DG methods mainly focusing on the loss function design. In this paper, we propose to explore an orthogonal direction, i.e., the design of the backbone architecture. It is motivated by an empirical finding that transformer-based models trained with empirical risk minimization (ERM) outperform CNN-based models employing state-of-the-art (SOTA) DG algorithms on multiple DG datasets. We develop a formal framework to characterize a network's robustness to distribution shifts by studying its architecture's alignment with the correlations in the dataset. This analysis guides us to propose a novel DG model built upon vision transformers, namely \emph{Generalizable Mixture-of-Experts (GMoE)}. Extensive experiments on DomainBed demonstrate that GMoE trained with ERM outperforms SOTA DG baselines by a large margin. Moreover, GMoE is complementary to existing DG methods and its performance is substantially improved when trained with DG algorithms.
****
PEER: A Collaborative Language Model
Timo Schick,Jane A. Yu,Zhengbao Jiang,Fabio Petroni,Patrick Lewis,Gautier Izacard,Qingfei You,Christoforos Nalmpantis,Edouard Grave,Sebastian Riedel
Textual content is often the output of a collaborative writing process: We start with an initial draft, ask for suggestions, and repeatedly make changes.
Agnostic of this process, today’s language models are trained to generate only the final result. As a consequence, they lack several abilities crucial for collaborative writing: They are unable to update existing texts, difficult to control and incapable of verbally planning or explaining their actions.
To address these shortcomings, we introduce PEER, a collaborative language model that is trained to imitate the entire writing process itself. PEER can write drafts, add suggestions, propose edits and provide explanations for its actions. Crucially, we train multiple instances of PEER able to infill various parts of the writing process, enabling the use of self-training techniques for increasing the quality, amount and diversity of training data. This unlocks PEER's full potential by making it applicable in domains for which no edit histories are available and improving its ability to follow instructions, to write useful comments, and to explain its actions. We show that PEER achieves strong performance across various domains and editing tasks.
****
A simple but effective and efficient global modeling paradigm for image restoration
man zhou,Jie Huang,Jie Xiao,Hu Yu,Danfeng Hong,Chongyi Li
Global modelling-based image restoration frameworks (e.g., Transformer-like arch

itecture) has gained popularity. Despite the remarkable advancement, the success may be at the cost of model parameters and FLOPs while the intrinsic characteristics of specific task are ignored. The objective of our work is orthogonal to previous studies and we thus tailor a simple yet effective global modelling paradigm for image restoration. The key insights which motivate our study are two-fold: 1) Fourier transform is capable of disentangling image degradation and content component, acting as the image degradation prior embedded into image restoration framework; 2) Fourier domain innately embraces global property where each pixel of Fourier space is involved with all the spatial pixels. We obey the de facto global modeling rule ``spatial interaction + channel evolution" of previous studies. Differently, we customize the core designs: multi-scale Fourier period spatial modeling and Fourier channel evolution. Equipped with above designs, our image restoration paradigm is verified on mainstream image restoration tasks including image de-raining, image enhancement, image de-hazing, and guided image super-resolution. The extensive experiments suggest that our paradigm achieves the competitive performance with fewer computational resources. Our main focus is not to beat previous frameworks but hopes to provide an alternative global modelling-based customized image restoration framework. Code will be publicly available.
****

Empowering Networks With Scale and Rotation Equivariance Using A Similarity Convolution
Zikai Sun,Thierry Blu
The translational equivariant nature of Convolutional Neural Networks (CNNs) is a reason for its great success in computer vision. However, networks do not enjoy more general equivariance properties such as rotation or scaling, ultimately limiting their generalization performance. To address this limitation, we devise a method that endows CNNs with simultaneous equivariance with respect to translation, rotation, and scaling. Our approach defines a convolution-like operation and ensures equivariance based on our proposed scalable Fourier-Argand representation. The method maintains similar efficiency as a traditional network and hardly introduces any additional learnable parameters, since it does not face the computational issue that often occurs in group-convolution operators. We validate the efficacy of our approach in the image classification task, demonstrating its robustness and the generalization ability to both scaled and rotated inputs.
****

ISS: Image as Stepping Stone for Text-Guided 3D Shape Generation
Zhengzhe Liu,Peng Dai,Ruihui Li,XIAOJUAN QI,Chi-Wing Fu
Text-guided 3D shape generation remains challenging due to the absence of large paired text-shape dataset, the substantial semantic gap between these two modalities, and the structural complexity of 3D shapes. This paper presents a new framework called Image as Stepping Stone (ISS) for the task by introducing 2D image as a stepping stone to connect the two modalities and to eliminate the need for paired text-shape data. Our key contribution is a two-stage feature-space-alignment approach that maps CLIP features to shapes by harnessing a pre-trained single-view reconstruction (SVR) model with multi-view supervisions: first map the CLIP image feature to the detail-rich shape space in the SVR model, then map the CLIP text feature to the shape space and optimize the mapping by encouraging CLIP consistency between the input text and the rendered images. Further, we formulate a textguided shape stylization module to dress up the output shapes with novel structures and textures. Beyond existing works on 3D shape generation from text, our new approach is general for creating shapes in a broad range of categories, without requiring paired text-shape data. Experimental results manifest that our approach outperforms the state-of-the-arts and our baselines in terms of fidelity and consistency with text. Further, our approach can stylize the generated shapes with both realistic and fantasy structures and textures. Codes are available at https://github.com/liuzhengzhe/ISS-Image-as-Stepping-Stone-for-Text-Guided-3D-Shape-Generation.
****

Robust and Controllable Object-Centric Learning through Energy-based Models

Ruixiang ZHANG,Tong Che,Boris Ivanovic,Renhao Wang,Marco Pavone,Yoshua Bengio,Liam Paull

Humans are remarkably good at understanding and reasoning about complex visual scenes. The capability of decomposing low-level observations into discrete objects allows us to build a grounded abstract representation and identify the compositional structure of the world. Thus it is a crucial step for machine learning models to be capable of inferring objects and their properties from visual scene without explicit supervision. However, existing works on object-centric representation learning are either relying on tailor-made neural network modules or assuming sophisticated models of underlying generative and inference processes. In this work, we present EGO, a conceptually simple and general approach to learning object-centric representation through energy-based model. By forming a permutation-invariant energy function using vanilla attention blocks that are readily available in Transformers, we can infer object-centric latent variables via gradient-based MCMC methods where permutation equivariance is automatically guaranteed. We show that EGO can be easily integrated into existing architectures, and can effectively extract high-quality object-centric representations, leading to better segmentation accuracy and competitive downstream task performance. We empirically evaluate the robustness of the learned representation from EGO against distribution shift. Finally, we demonstrate the effectiveness of EGO in systematic compositional generalization, by recomposing learned energy functions for novel scene generation and manipulation.

****

Learning Antidote Data to Individual Unfairness

Peizhao Li,Ethan Xia,Hongfu Liu

Fairness is an essential factor for machine learning systems deployed in high-stake applications. Among all fairness notions, individual fairness, following a consensus that `similar individuals should be treated similarly,' is a vital notion to guarantee fair treatment for individual cases. Previous methods typically characterize individual fairness as a prediction-invariant problem when perturbing sensitive attributes, and solve it by adopting the Distributionally Robust Optimization (DRO) paradigm. However, adversarial perturbations along a direction covering sensitive information do not consider the inherent feature correlations or innate data constraints, and thus mislead the model to optimize at off-manifold and unrealistic samples. In light of this, we propose a method to learn and generate antidote data that approximately follows the data distribution to remedy individual unfairness. These on-manifold antidote data can be used through a generic optimization procedure with original training data, resulting in a pure pre-processing approach to individual unfairness, or can also fit well with the in-processing DRO paradigm. Through extensive experiments, we demonstrate our antidote data resists individual unfairness at a minimal or zero cost to the model's predictive utility.

****

Does Continual Learning Equally Forget All Parameters?

Haiyan Zhao,Tianyi Zhou,Guodong Long,Jing Jiang,Chengqi Zhang

Distribution shift (e.g., task or domain shift) in continual learning (CL) usually results in catastrophic forgetting of neural networks. Although it can be alleviated by repeatedly replaying buffered data, the every-step replay is time-consuming and the memory to store historical data is usually too small for retraining all parameters. In this paper, we study which modules in neural networks are more prone to forgetting by investigating their training dynamics during CL. Our proposed metrics show that only a few modules are more task-specific and sensitively alters between tasks, while others can be shared across tasks as common knowledge. Hence, we attribute forgetting mainly to the former and find that finetuning them only on a small buffer at the end of any CL method can bring non-trivial improvement. Due to the small number of finetuned parameters, such ``Forgetting Prioritized Finetuning (FPF)'' is efficient on both the computation and buffer size required. We further propose a more efficient and simpler method that entirely removes the every-step replay and replaces them by only $k$-times of FPF periodically triggered during CL. Surprisingly, this ``k-FPF'' performs compar ably to FPF and outperforms the SOTA CL methods but significantly reduces their computational overhead and cost. In experiments on several benchmarks of class- and domain-incremental CL, FPF consistently improves existing CL methods by a large margin and k-FPF further excels on the efficiency without degrading the accuracy. We also empirically studied the impact of buffer size, epochs per task, and finetuning modules to the cost and accuracy of our methods.

****

STREET: A MULTI-TASK STRUCTURED REASONING AND EXPLANATION BENCHMARK

Danilo Neves Ribeiro,Shen Wang,Xiaofei Ma,Henghui Zhu,Rui Dong,Deguang Kong,Juliette Burger,Anjelica Ramos,zhiheng huang,William Yang Wang,George Karypis,Bing Xiang,Dan Roth

We introduce STREET, a unified multi-task and multi-domain natural language reasoning and explanation benchmark. Unlike most existing question-answering (QA) datasets, we expect models to not only answer questions, but also produce step-by-step structured explanations describing how premises in the question are used to produce intermediate conclusions that can prove the correctness of a certain answer. We perform extensive evaluation with popular language models such as few-shot prompting GPT-3 and fine-tuned T5. We find that these models still lag behind human performance when producing such structured reasoning steps. We believe this work will provide a way for the community to better train and test systems on multi-step reasoning and explanations in natural language.

****

Topology-aware Robust Optimization for Out-of-Distribution Generalization

Fengchun Qiao,Xi Peng

Out-of-distribution (OOD) generalization is a challenging machine learning problem yet highly desirable in many high-stake applications.

Existing methods suffer from overly pessimistic modeling with low generalization confidence. As generalizing to arbitrary test distributions is impossible, we hypothesize that further structure on the topology of distributions is crucial in developing strong OOD resilience. To this end, we propose topology-aware robust optimization (TRO) that seamlessly integrates distributional topology in a principled optimization framework. More specifically, TRO solves two optimization objectives: (1) Topology Learning which explores data manifold to uncover the distributional topology; (2) Learning on Topology which exploits the topology to constrain robust optimization for tightly-bounded generalization risks. We theoretically demonstrate the effectiveness of our approach, and empirically show that it significantly outperforms the state of the arts in a wide range of tasks including classification, regression, and semantic segmentation. Moreover, we empirically find the data-driven distributional topology is consistent with domain knowledge, enhancing the explainability of our approach.

****

Exploring Neural Network Representational Similarity using Filter Subspaces

Wei Chen,Zichen Miao,Qiang Qiu

Analyzing representational similarity in neural networks is crucial to numerous tasks, such as interpreting or transferring deep models. One typical approach is to input probing data into convolutional neural networks (CNNs) as stimuli to reveal their deep representation for model similarity analysis. Those methods are often computationally expensive and stimulus-dependent. By representing filter subspace in a CNN as a set of filter atoms, previous work has reported competitive performance in continual learning by learning a different set of filter atoms for each task while sharing common atom coefficients across tasks. Inspired by this observation, in this paper, we propose a new paradigm for reducing representational similarity analysis in CNNs to filter subspace distance assessment. Specifically, when filter atom coefficients are shared across networks, model representational similarity can be significantly simplified as calculating the cosine distance among respective filter atoms, to achieve \textit{millions of times} computation reduction. We provide both theoretical and empirical evidence that this simplified filter subspace-based similarity preserves a strong linear correlation with other popular stimulus-based metrics, while being significantly more efficient and robust to probing data. We further validate the effectiveness of th e proposed method in various applications, such as analyzing training dynamics as well as in federated and continual learning. We hope our findings can help further explorations of real-time large-scale representational similarity analysis in neural networks.

****

EAGLE: Large-scale Learning of Turbulent Fluid Dynamics with Mesh Transformers

Steeven JANNY,Aurélien Bénéteau,Madiha Nadri,Julie Digne,Nicolas THOME,Christian Wolf

Estimating fluid dynamics is classically done through the simulation and integration of numerical models solving the Navier-Stokes equations, which is computationally complex and time-consuming even on high-end hardware. This is a notoriously hard problem to solve, which has recently been addressed with machine learning, in particular graph neural networks (GNN) and variants trained and evaluated on datasets of static objects in static scenes with fixed geometry. We attempt to go beyond existing work in complexity and introduce a new model, method and benchmark. We propose EAGLE: a large-scale dataset of ~1.1 million 2D meshes resulting from simulations of unsteady fluid dynamics caused by a moving flow source interacting with nonlinear scene structure of varying geometries, with 600 different scenes of three different types in total. To perform future forecasting of pressure and velocity on the challenging EAGLE dataset, we introduce a new mesh transformer. It leverages node clustering, graph pooling and global attention to learn long-range dependencies between spatially distant data points without needing a large number of iterations, as existing GNN methods do. We show that our transformer outperforms state-of-the-art performance on, both, existing synthetic and real datasets and on EAGLE. Finally, we highlight that our approach learns to attend to airflow, integrating complex information in a single iteration.

****

Limitless Stability for Graph Convolutional Networks

Christian Koke

This work establishes rigorous, novel and widely applicable stability guarantees and transferability bounds for general graph convolutional networks -- without reference to any underlying limit object or statistical distribution. Crucially, utilized graph-shift operators are not necessarily assumed to be normal, allowing for the treatment of networks on both directed- and undirected graphs within the developed framework. In the undirected setting, stability to node-level perturbations is related to an 'adequate spectral covering' property of the filters in each layer. Stability to edge-level perturbations is discussed and related to properties of the utilized filters such as their Lipschitz constants. Results on stability to vertex-set non-preserving perturbations are obtained by utilizing recently developed mathematical-physics based tools. As an exemplifying application of the developed theory, it is showcased that
general graph convolutional networks utilizing the un-normalized graph Laplacian as graph-shift-operator can be rendered stable to collapsing strong edges in the underlying graph if filters are mandated to be constant at infinity. These theoretical results are supported by corresponding numerical investigations showcasing the response of filters and networks to such perturbations.

****

MiSAL: Active Learning for Every Budget

Guy Hacohen,Daphna Weinshall

In supervised Active Learning (AL), the learner can manipulate the labeled training set by choosing examples to be labeled by an oracle. The size of the labeled set is termed budget. Recent years have seen significant progress in this domain in the context of deep active learning. In particular, it has been shown that in general, different families of AL strategies are suitable for high and low budgets. Here we address for the first time the problem of deciding which family of strategies is most suitable for a given budget in a given task. We start from the theoretical analysis of a mixture model, which motivates a computational approach to decide on the most suitable family of methods for the task and budget at hand. We then propose a practical decision algorithm, which determines what family of strategies should be preferred. Using this algorithm, we introduce MiSAL

- a mixed strategy active learning algorithm. MiSAL combines AL strategies from different families, resulting in a method that fits all budgets. We support the analysis by an empirical study, showing the superiority of our method when dealing with image datasets.
****

SOM-CPC: Unsupervised Contrastive Learning with Self-Organizing Maps for Structured Representations of High-Rate Time Series
Iris A.M. Huijben,Arthur Andreas Nijdam,Sebastiaan Overeem,Merel M Van Gilst,Ruud Van Sloun
Continuous monitoring with an ever-increasing number of sensors has become ubiquitous across many application domains. Acquired data are typically high-dimensional and difficult to interpret, but they are also hypothesized to lie on a low-dimensional manifold. Dimensionality reduction techniques have, therefore, been sought for. Popular linear methods like Principle Component Analysis (PCA) have been extended to non-linear techniques such as Self-Organizing Maps (SOMs) or deep learning (DL) models. DL models have the ability to act on raw data, preventing heuristic feature selection, but the resulting latent space is often unstructured and still multi-dimensional. PCA and SOMs, on the other hand, need to be preceded with a feature-extraction step, but can then map high-dimensional features to 2D space. In this work we propose SOM-CPC, a model that jointly optimizes Contrastive Predictive Coding and a SOM to find an organized 2D manifold. We address a largely unexplored and challenging set of scenarios comprising high-rate time series, and show on both synthetic and real-life data (medical sleep data and audio recordings) that SOM-CPC outperforms both DL-based feature extraction, followed by PCA, K-means or a SOM, and strong deep-SOM baselines that jointly optimize a DL model and a SOM. SOM-CPC has great potential to expose latent patterns in high-rate data streams and may therefore contribute to a better understanding of many different processes and systems.
****

DIVISION: Memory Efficient Training via Dual Activation Precision
Guanchu Wang,Zirui Liu,Zhimeng Jiang,Ninghao Liu,na Zou,Xia Hu
Activation compressed training (ACT) has been shown to be a promising way to reduce the memory cost of training deep neural networks (DNNs). However, existing work of ACT relies on searching for optimal bit-width during DNN training to reduce the quantization noise, which makes the procedure complicated and less transparent. To this end, we propose a simple and effective method to compress DNN training. Our method is motivated by an instructive observation: DNN backward propagation mainly utilizes the low-frequency component (LFC) of the activation maps, while the majority of memory is for caching the high-frequency component (HFC) during the training. This indicates the HFC of activation maps is highly redundant and compressible during DNN training, which inspires our proposed Dual Activation Precision (DIVISION). During the training, DIVISION preserves the high-precision copy of LFC and compresses the HFC into a light-weight copy with low numerical precision. This can significantly reduce the memory cost without negatively affecting the precision of backward propagation such that DIVISION maintains competitive model accuracy. Experimental results show DIVISION achieves over 10× compression of activation maps, and significantly higher training throughput than state-of-the-art ACT methods, without loss of model accuracy. The code is available at https://anonymous.4open.science/r/division-5CC0/

****

CLIP-PAE: Projection-Augmentation Embedding to Extract Relevant Features for a Disentangled, Interpretable and Controllable Text-Guided Image Manipulation
Chenliang Zhou,Fangcheng Zhong,Cengiz Oztireli
Recently introduced Contrastive Language-Image Pre-Training (CLIP) bridges images and text by embedding them into a joint latent space. This opens the door to ample literature that aims to manipulate an input image by providing a textual explanation. However, due to the discrepancy between image and text embeddings in the joint space, using text embeddings as the optimization target often introduces undesired artifacts in the resulting images. Disentanglement, interpretabilit

y, and controllability are also hard to guarantee for manipulation. To alleviate these problems, we propose to define corpus subspaces spanned by prompts to capture specific image characteristics. We introduce CLIP projection-augmentation embedding (PAE) as an optimization target to improve the performance of text-guided image manipulation. Our method is a simple and general paradigm that can be easily computed and adapted, and smoothly incorporated into any CLIP-based latent manipulation algorithm to improve performance. To demonstrate the effectiveness of our method, we conduct several theoretical and empirical system studies. As a case study, we utilize the method for text-guided semantic face editing. We quantitatively and qualitatively demonstrate that PAE facilitates a more disentangled, interpretable, and controllable image manipulation method with state of the art quality and accuracy.

****

Token Merging: Your ViT But Faster

Daniel Bolya,Cheng-Yang Fu,Xiaoliang Dai,Peizhao Zhang,Christoph Feichtenhofer,Judy Hoffman

We introduce Token Merging (ToMe), a simple method to increase the throughput of existing ViT models without needing to train. ToMe gradually combines similar tokens in a transformer using a general and light-weight matching algorithm that is as fast as pruning while being more accurate. Off-the-shelf, ToMe can 2x the throughput of state-of-the-art ViT-L @ 512 and ViT-H @ 518 models on images and 2.2x the throughput of ViT-L on video with only a 0.2-0.3% accuracy drop in each case. ToMe can also easily be applied during training, improving in practice training speed up to 2x for MAE fine-tuning on video. Training with ToMe further minimizes accuracy drop, leading to 2x the throughput of ViT-B on audio for only a 0.4% mAP drop. Qualitatively, we find that ToMe merges object parts into one token, even over multiple frames of video. Overall, ToMe’s accuracy and speed are competitive with state-of-the-art on images, video, and audio.

****

Provable Adaptivity in Adam

Bohan Wang,Yushun Zhang,Huishuai Zhang,Qi Meng,Zhi-Ming Ma,Tie-Yan Liu,Wei Chen

Adaptive Moment Estimation (Adam) has been observed to converge faster than stochastic gradient descent (SGD) in practice. However, such an advantage has not been theoretically characterized -- the existing convergence rate of Adam is no better than SGD. We attribute this mismatch between theory and practice to a commonly used assumption: the gradient is globally Lipschitz continuous (called L-smooth condition). Specifically, compared to SGD, Adam adaptively chooses a learning rate better suited to the local gradient Lipschitz constant (called local smoothness). This effect becomes
prominent when the local smoothness varies drastically across the domain.

In this paper, we analyze the convergence of Adam under a condition called (L_0,L_1)-smooth condition, which allows the gradient Lipschitz constant to change with the gradient norm. This condition has been empirically verified to be more realistic for deep neural networks \citep{zhang2019gradient} than the L-smooth condition. Under (L_0,L_1)-smooth condition, we establish the convergence for Adam with practical hyperparameters. As such, we argue that Adam can adapt to this local smoothness condition, justifying Adam's \emph{adaptivity}. In contrast, SGD can be arbitrarily slow under this condition.   Our result can shed light on the benefit of adaptive gradient methods over non-adaptive ones.

****

De Novo Molecular Generation via Connection-aware Motif Mining

Zijie Geng,Shufang Xie,Yingce Xia,Lijun Wu,Tao Qin,Jie Wang,Yongdong Zhang,Feng Wu,Tie-Yan Liu

De novo molecular generation is an essential task for science discovery. Recently, fragment-based deep generative models have attracted much research attention due to their flexibility in generating novel molecules based on existing molecule fragments. However, the motif vocabulary, i.e., the collection of frequent fragments, is usually built upon heuristic rules, which brings difficulties to capturing common substructures from large amounts of molecules. In this work, we propose MiCaM to generate molecules based on mined connection-aware motifs. Specifi cally, it leverages a data-driven algorithm to automatically discover motifs from a molecule library by iteratively merging subgraphs based on their frequency. The obtained motif vocabulary consists of not only molecular motifs (i.e., the frequent fragments), but also their connection information, indicating how the motifs are connected with each other. Based on the mined connection-aware motifs, MiCaM builds a connection-aware generator, which simultaneously picks up motifs and determines how they are connected. We test our method on distribution-learning benchmarks (i.e., generating novel molecules to resemble the distribution of a given training set) and goal-directed benchmarks (i.e., generating molecules with target properties), and achieve significant improvements over previous fragment-based baselines. Furthermore, we demonstrate that our method can effectively mine domain-specific motifs for different tasks.

****

GANet: Graph-Aware Network for Point Cloud Completion with Displacement-Aware Point Augmentor

Hongyu Yan,Haoxi Ran,Li Lu

Remarkably, real-world data (e.g., LiDAR-based point clouds) is commonly sparse, uneven, occluded, and truncated. The point cloud completion task draws due attention, which aims to predict a complete and accurate shape from its partial observation. However, existing methods commonly adopt PointNet or PointNet++ to extract features of incomplete point clouds. In this paper, we propose an end-to-end Graph-Aware Network (\textbf{GANet}) to effectively learn from the contour information of the partial point clouds. Moreover, we design Displacements-Aware Augmentor (DPA) to upsample and refine coarse point clouds. With our graph-based feature extractors and Displacements-Aware Transformer, our DPA can precisely capture the geometric and structural features to refine the complete point clouds. Experiments on PCN and MVP datasets demonstrate that our GANet achieves state-of-the-art on the task of shape completion.

****

Multiple output samples for each input in a single-output Gaussian process

Jeremy Wong,Huayun Zhang,Nancy F. Chen

The standard Gaussian Process (GP) is formulated to only consider a single output sample for each input in the training set. Datasets for subjective tasks, such as spoken language assessment, may be annotated with output labels from multiple human raters for each input. This paper proposes to generalise the GP to allow for multiple output samples per input in the training set. This differs from a multi-output GP, because all output samples are from the same task here. The output density function is formulated to be the joint likelihood of observing all output samples. Through this, the hyper-parameters are optimised using a criterion that is similar to minimising a Kullback-Leibler divergence. The test set predictions are inferred fairly similarly to a standard GP, with a key difference being in the optimised hyper-parameters. This approach is evaluated on spoken language assessment tasks, using the public speechocean762 dataset and an internal Tamil language dataset. The results show that by using the proposed method, the GP is able to compute a test set output distribution that is more similar to the collection of reference outputs annotated by multiple human raters.

****

Demystifying the Optimization and Generalization of Deep PAC-Bayesian Learning

Wei Huang,Chunrui Liu,Yilan Chen,Richard Yi Da Xu,Miao Zhang,Tsui-Wei Weng

In addition to being a successful generalization bound analysis tool, the PAC-Bayesian bound can also be incorporated into an objective function to train a probabilistic neural network, which we refer to simply as {\it PAC-Bayesian Learning}. PAC-Bayesian learning has been proven to be able to achieve a competitive expected test set error numerically, while providing a tight generalization bound in practice, through gradient descent training. Despite its empirical success, the theoretical analysis of deep PAC-Bayesian learning for neural networks is rarely explored. To this end, this paper proposes a theoretical convergence and generalization analysis for PAC-Bayesian learning. For a deep and wide probabilistic neural network, we show that when PAC-Bayesian learning is applied, the convergence result corresponds to solving a kernel ridge regression when the probabilis tic neural tangent kernel (PNTK) is used as its kernel. Based on this finding, we further obtain an analytic and guaranteed PAC-Bayesian generalization bound for the first time, which is an improvement over the Rademacher complexity-based bound for deterministic neural networks. Finally, drawing insight from our theoretical results, we propose a proxy measure for efficient hyperparameter selection, which is proven to be time-saving on various benchmarks.

****

Revisiting the Entropy Semiring for Neural Speech Recognition

Oscar Chang,Dongseong Hwang,Olivier Siohan

In streaming settings, speech recognition models have to map sub-sequences of speech to text before the full audio stream becomes available. However, since alignment information between speech and text is rarely available during training, models need to learn it in a completely self-supervised way. In practice, the exponential number of possible alignments makes this extremely challenging, with models often learning peaky or sub-optimal alignments. Prima facie, the exponential nature of the alignment space makes it difficult to even quantify the uncertainty of a model's alignment distribution. Fortunately, it has been known for decades that the entropy of a probabilistic finite state transducer can be computed in time linear to the size of the transducer via a dynamic programming reduction based on semirings. In this work, we revisit the entropy semiring for neural speech recognition models, and show how alignment entropy can be used to supervise models through regularization or distillation. We also contribute an open-source implementation of CTC and RNN-T in the semiring framework that includes numerically stable and highly parallel variants of the entropy semiring. Empirically, we observe that the addition of alignment distillation improves the accuracy and latency of an already well-optimized teacher-student distillation model, achieving state-of-the-art performance on the Librispeech dataset in the streaming scenario.

****

Rethinking skip connection model as a learnable Markov chain

Chen Dengsheng,Jie Hu,Wenwen Qiang,Xiaoming Wei,Enhua Wu

Over the past few years afterward the birth of ResNet, skip connection has become the defacto standard for the design of modern architectures due to its widespread adoption, easy optimization, and proven performance.

Prior work has explained the effectiveness of the skip connection mechanism from different perspectives.

In this work, we deep dive into the model's behaviors with skip connections which can be formulated as a learnable Markov chain.

An efficient Markov chain is preferred as it always maps the input data to the target domain in a better way.

However, while a model is explained as a Markov chain, it is not guaranteed to be optimized following an efficient Markov chain by existing SGD-based optimizers prone to getting trapped in local optimal points.

In order to move towards a more efficient Markov chain, we propose a simple routine of penal connection to make any residual-like model become a learnable Markov chain.

Aside from that, the penal connection can also be viewed as a particular model regularization and can be easily implemented with one line of code in the most popular deep learning frameworks.

The encouraging experimental results in multi-modal translation and image recognition empirically confirm our conjecture of the learnable Markov chain view and demonstrate the superiority of the proposed penal connection.

****

Activation Function: Absolute Function,One Function Behaves more Individualized

Jinxin Wei,Zhe Hou

Inspire by nature world mode, a activation function is proposed. It is absolute function.Through test on mnist dataset and fully-connected neural network and convolutional neural network, some conclusions are put forward. The line of accuracy of absolute function is a little shaken that is different from the line of accuracy of relu and leaky relu. The absolute function can keep the negative parts as equal as the positive parts, so the individualization is more active than relu and leaky relu function. The absolute function is less likely to be over-fitting. Through test on mnist and autoencoder, It is that the leaky relu function can do classification task well, while the absolute function can do generation task well. Because the classification task need more universality and generation task need more individualization. The pleasure irritation and painful irritation is not only the magnitude differences, but also the sign differences, so the negative parts should keep as a part. Stimulation which happens frequently is low value, it is showed around zero in figure 1 . Stimulation which happens accidentally is high value, it is showed far away from zero in figure 1. So the high value is the big stimulation, which is individualization.

****

Measuring axiomatic soundness of counterfactual image models

Miguel Monteiro,Fabio De Sousa Ribeiro,Nick Pawlowski,Daniel C. Castro,Ben Glocker

We present a general framework for evaluating image counterfactuals. The power and flexibility of deep generative models make them valuable tools for learning mechanisms in structural causal models. However, their flexibility makes counterfactual identifiability impossible in the general case.

Motivated by these issues, we revisit Pearl's axiomatic definition of counterfactuals to determine the necessary constraints of any counterfactual inference model: composition, reversibility, and effectiveness. We frame counterfactuals as functions of an input variable, its parents, and counterfactual parents and use the axiomatic constraints to restrict the set of functions that could represent the counterfactual, thus deriving distance metrics between the approximate and ideal functions. We demonstrate how these metrics can be used to compare and choose between different approximate counterfactual inference models and to provide insight into a model's shortcomings and trade-offs.

****

Alternating Differentiation for Optimization Layers

Haixiang Sun,Ye Shi,Jingya Wang,Hoang Duong Tuan,H. Vincent Poor,Dacheng Tao

The idea of embedding optimization problems into deep neural networks as optimization layers to encode constraints and inductive priors has taken hold in recent years. Most existing methods focus on implicitly differentiating Karush–Kuhn–Tucker (KKT) conditions in a way that requires expensive computations on the Jacobian matrix, which can be slow and memory-intensive. In this paper, we developed a new framework, named Alternating Differentiation (Alt-Diff), that differentiates optimization problems (here, specifically in the form of convex optimization problems with polyhedral constraints) in a fast and recursive way. Alt-Diff decouples the differentiation procedure into a primal update and a dual update in an alternating way. Accordingly, Alt-Diff substantially decreases the dimensions of the Jacobian matrix especially for optimization with large-scale constraints and thus increases the computational speed of implicit differentiation. We show that the gradients obtained by Alt-Diff are consistent with those obtained by differentiating KKT conditions. In addition, we propose to truncate Alt-Diff to further accelerate the computational speed. Under some standard assumptions, we show that the truncation error of gradients is upper bounded by the same order of variables' estimation error. Therefore, Alt-Diff can be truncated to further increase computational speed without sacrificing much accuracy. A series of comprehensive experiments validate the superiority of Alt-Diff.

****

Out-of-distribution Detection with Implicit Outlier Transformation

Qizhou Wang,Junjie Ye,Feng Liu,Quanyu Dai,Marcus Kalander,Tongliang Liu,Jianye HAO,Bo Han

Outlier exposure (OE) is powerful in out-of-distribution (OOD) detection, enhancing detection capability via model fine-tuning with surrogate OOD data. However, surrogate data typically deviate from test OOD data. Thus, the performance of OE when facing unseen OOD data, can be weaken. To address this issue, we propose a novel OE-based approach that makes the model perform well for unseen OOD situations, even for unseen OOD cases. It leads to a min-max learning scheme---search ing to synthesize OOD data that leads to worst judgments and learning from such OOD data for the uniform performance in OOD detection. In our realization, these worst OOD data are synthesized by transforming original surrogate ones, where the associated transform functions are learned implicitly based on our novel insight that model perturbation leads to data transformation. Our methodology offers an efficient way of synthesizing OOD data, which can further benefit the detection model, besides the surrogate OOD data. We conduct extensive experiments under various OOD detection setups, demonstrating the effectiveness of our method against its advanced counterparts.

****

Parameter Averaging for Feature Ranking

Talip Ucar,Ehsan Hajiramezanali

Neural Networks are known to be sensitive to initialisation. The methods that rely on neural networks for feature ranking are not robust since they can have variations in their ranking when the model is initialized and trained with different random seeds. In this work, we introduce a novel method based on parameter averaging to estimate accurate and robust feature importance in tabular data setting, referred as XTab. We first initialize and train multiple instances of a shallow network (referred as local masks) with "different random seeds" for a downstream task. We then obtain a global mask model by "averaging the parameters" of local masks. We show that although the parameter averaging might result in a global model with higher loss, it still leads to the discovery of the ground-truth feature importance more consistently than an individual model does. We conduct extensive experiments on a variety of synthetic and real-world data, demonstrating that the XTab can be used to obtain the global feature importance that is not sensitive to sub-optimal model initialisation.

****

Gradient Estimation for Unseen Domain Risk Minimization with Pre-Trained Models

Byounggyu Lew,Donghyun Son,Buru Chang

Domain generalization aims to build generalized models that perform well on unseen domains when only source domains are available for model optimization. Recent studies have demonstrated that large-scale pre-trained models could play an important role in domain generalization by providing their generalization power. However, large-scale pre-trained models are not fully equipped with target task-specific knowledge due to a discrepancy between the pre-training objective and the target task. Although the task-specific knowledge could be learned from source domains by fine-tuning, this hurts the generalization power of the pre-trained models because of gradient bias toward the source domains. To address this issue, we propose a new domain generalization method that estimates unobservable gradients that reduce potential risks in unseen domains, using a large-scale pre-trained model. Our proposed method allows the pre-trained model to learn task-specific knowledge further while preserving its generalization ability with the estimated gradients. Experimental results show that our proposed method outperforms baseline methods on DomainBed, a standard benchmark in domain generalization. We also provide extensive analyses to demonstrate that the estimated unobserved gradients relieve the gradient bias, and the pre-trained model learns the task-specific knowledge without sacrificing its generalization power.

****

Nearing or Surpassing: Overall Evaluation of Human-Machine Dynamic Vision Ability

Shiyu Hu,Xin Zhao,Yipei Wang,Yanhu Shan,Kaiqi Huang

Dynamic visual ability (DVA), a fundamental function of the human visual system, has been successfully modeled by many computer vision tasks in recent decades. However, the prosperity developments mainly concentrate on using deep neural networks (DNN) to simulate the human DVA system, but evaluation systems still simply compare performance between machines, making it tough to determine how far the gap is between humans and machines in dynamic vision tasks. In fact, neglecting this issue not only makes it hard to determine the correctness of current research routes, but also cannot truly measure the DVA intelligence of machines. To answer the question, this work designs a comprehensive evaluation system based on the 3E paradigm -- we carefully pick 87 videos from various dimensions to construct the environment, confirming it can cover both perceptual and cognitive components of DVA; select 20 representative machines and 15 human subjects to form the task executors, ensuring that different model structures can help us observe the effectiveness of research development; and finally quantify their DVA with a strict evaluation process. Based on detailed experimental analyses, we first determine that the current algorithm research route has effectively shortened the gap. Besides, we further summarize the weaknesses of different executors, and design a human-machine cooperation mechanism with superhuman performance. In summary, the contributions include: (1) Quantifying the DVA of humans and machines, (2) proposing a new view to evaluate DVA intelligence based on the human-machine comparison, and (3) providing a possibility of human-machine cooperation. The datasets, toolkits, codes, and evaluation metrics will be open-sourced to help researchers develop intelligent research on dynamic vision tasks.

****

Extracting Robust Models with Uncertain Examples

Guanlin Li,Guowen Xu,Shangwei Guo,Han Qiu,Jiwei Li,Tianwei Zhang

Model extraction attacks are proven to be a severe privacy threat to Machine Learning as a Service (MLaaS). A variety of techniques have been designed to steal a remote machine learning model with high accuracy and fidelity. However, how to extract a robust model with similar resilience against adversarial attacks is never investigated. This paper presents the first study toward this goal. We first analyze those existing extraction solutions either fail to maintain the model accuracy or model robustness or lead to the robust overfitting issue. Then we propose Boundary Entropy Searching Thief (BEST), a novel model extraction attack to achieve both accuracy and robustness extraction under restricted attack budgets. BEST generates a new kind of uncertain examples for querying and reconstructing the victim model. These samples have uniform confidence scores across different classes, which can perfectly balance the trade-off between model accuracy and robustness. Extensive experiments demonstrate that BEST outperforms existing attack methods over different datasets and model architectures under limited data. It can also effectively invalidate state-of-the-art extraction defenses.

****

Neural Groundplans: Persistent Neural Scene Representations from a Single Image

Prafull Sharma,Ayush Tewari,Yilun Du,Sergey Zakharov,Rares Andrei Ambrus,Adrien Gaidon,William T. Freeman,Fredo Durand,Joshua B. Tenenbaum,Vincent Sitzmann

We present a method to map 2D image observations of a scene to a persistent 3D scene representation, enabling novel view synthesis and disentangled representation of the movable and immovable components of the scene. Motivated by the bird’s-eye-view (BEV) representation commonly used in vision and robotics, we propose conditional neural groundplans, ground-aligned 2D feature grids, as persistent and memory-efficient scene representations. Our method is trained self-supervised from unlabeled multi-view observations using differentiable rendering, and learns to complete geometry and appearance of occluded regions. In addition, we show that we can leverage multi-view videos at training time to learn to separately reconstruct static and movable components of the scene from a single image at test time. The ability to separately reconstruct movable objects enables a variety of downstream tasks using simple heuristics, such as extraction of object-centric 3D representations, novel view synthesis, instance-level segmentation, 3D bounding box prediction, and scene editing. This highlights the value of neural groundplans as a backbone for efficient 3D scene understanding models.

****

Semi-supervised Counting via Pixel-by-pixel Density Distribution Modelling

Hui LIN,Zhiheng Ma,Rongrong Ji,Yaowei Wang,su zhou,Xiaopeng Hong

This paper focuses on semi-supervised crowd counting, where only a small portion of the training data are labeled. We formulate the pixel-wise density value to regress as a probability distribution, instead of a single deterministic value, and utilize a dual-branch structure to model the corresponding discrete form of the distribution function. On the basis, we propose a semi-supervised crowd counting model. Firstly, we enhance the transformer decoder by usingdensity tokens t o specialize the forwards of decoders w.r.t. different density intervals; Secondly, we design a pixel-wise distribution matching loss to measure the differences in the pixel-wise density distributions between the prediction and the ground-truth; Thirdly, we propose an interleaving consistency regularization term to align the prediction of two branches and make them consistent. Extensive experiments on four datasets are performed to show that our method clearly outperforms the competitors by a large margin under various labeled ratio settings.

****

Understanding Self-Supervised Pretraining with Part-Aware Representation Learning

Jiyang Qi,Jie Zhu,Mingyu Ding,Xiaokang Chen,Ping Luo,Leye Wang,Xinggang Wang,Wenyu Liu,Jingdong Wang

In this paper, we are interested in understanding self-supervised pretraining through studying the capability that self-supervised representation pretraining methods learn part-aware representations. The study is mainly motivated by that random views, used in contrastive learning, and random masked (visible) patches, used in masked image modeling, are often about object parts.

We explain that masked image modeling is a part-to-part task: the masked patches of the object are hallucinated from the visible patches, and that contrastive learning is a part-to-whole task: the projection layer hallucinates the whole object representation from the object part representation learned from the encoder. The explanation suggests that the self-supervised pretrained encoder is required to understand the object part. We empirically compare the off-the-shelf encoders pretrained with several representative methods on object-level recognition and part-level recognition. The results show that the fully-supervised model outperforms self-supervised models for object-level recognition, and most self-supervised contrastive learning and masked image modeling methods outperform the fully-supervised method for part-level recognition. It is observed that the combination of contrastive learning and masked image modeling further improves the performance.

****

E-CRF: Embedded Conditional Random Field for Boundary-caused Class Weights Confusion in Semantic Segmentation

Jie Zhu,Huabin Huang,Banghuai Li,Leye Wang

Modern semantic segmentation methods devote much effect to adjusting image feature representations to improve the segmentation performance in various ways, such as architecture design, attention mechnism, etc. However, almost all those methods neglect the particularity of class weights (in the classification layer) in segmentation models. In this paper, we notice that the class weights of categories that tend to share many adjacent boundary pixels lack discrimination, thereby limiting the performance. We call this issue Boundary-caused Class Weights Confusion (BCWC). We try to focus on this problem and propose a novel method named Embedded Conditional Random Field (E-CRF) to alleviate it. E-CRF innovatively fuses the CRF into the CNN network as an organic whole for more effective end-to-end optimization. The reasons are two folds. It utilizes CRF to guide the message passing between pixels in high-level features to purify the feature representation of boundary pixels, with the help of inner pixels belonging to the same object. More importantly, it enables optimizing class weights from both scale and direction during backpropagation. We make detailed theoretical analysis to prove it. Besides, superpixel is integrated into E-CRF and served as an auxiliary to exploit the local object prior for more reliable message passing. Finally, our proposed method yields impressive results on ADE20K, Cityscapes, and Pascal Context datasets.

****

Sample Complexity of Nonparametric Off-Policy Evaluation on Low-Dimensional Manifolds using Deep Networks

Xiang Ji,Minshuo Chen,Mengdi Wang,Tuo Zhao

We consider the off-policy evaluation problem of reinforcement learning using deep convolutional neural networks. We analyze the deep fitted Q-evaluation method for estimating the expected cumulative reward of a target policy, when the data are generated from an unknown behavior policy. We show that, by choosing network size appropriately, one can leverage any low-dimensional manifold structure in the Markov decision process and obtain a sample-efficient estimator without suffering from the curse of high data ambient dimensionality. Specifically, we establish a sharp error bound for fitted Q-evaluation, which depends on the intrinsic dimension of the state-action space, the smoothness of Bellman operator, and a function class-restricted χ^2-divergence. It is noteworthy that the restricted χ^2-divergence measures the behavior and target policies' {\it mismatch in the function space}, which can be small even if the two policies are not close to each other in their tabular forms. We also develop a novel approximation result for convolutional neural networks in Q-function estimation. Numerical experiments are provided to support our theoretical analysis.

****

Stochastic Differentially Private and Fair Learning
Andrew Lowy,Devansh Gupta,Meisam Razaviyayn
Machine learning models are increasingly used in high-stakes decision-making systems. In such applications, a major concern is that these models sometimes discriminate against certain demographic groups such as individuals with certain race, gender, or age. Another major concern in these applications is the violation of the privacy of users. While fair learning algorithms have been developed to mitigate discrimination issues, these algorithms can still leak sensitive information, such as individuals’ health or financial records. Utilizing the notion of differential privacy (DP), prior works aimed at developing learning algorithms that are both private and fair. However, existing algorithms for DP fair learning are either not guaranteed to converge or require full batch of data in each iteration of the algorithm to converge. In this paper, we provide the first stochastic differentially private algorithm for fair learning that is guaranteed to converge. Here, the term “stochastic" refers to the fact that our proposed algorithm converges even when minibatches of data are used at each iteration (i.e. stochastic optimization). Our framework is flexible enough to permit different fairness notions, including demographic parity and equalized odds. In addition, our algorithm can be applied to non-binary classification tasks with multiple (non-binary) sensitive attributes. As a byproduct of our convergence analysis, we provide the first utility guarantee for a DP algorithm for solving nonconvex-strongly concave min-max problems. Our numerical experiments show that the proposed algorithm consistently offers significant performance gains over the state-of-the-art baselines, and can be applied to larger scale problems with non-binary target/sensitive attributes.

****

CLIP-FLOW: CONTRASTIVE LEARNING WITH ITERATIVE PSEUDO LABELING FOR OPTICAL FLOW
Zhiqi Zhang,Nitin Bansal,Changjiang Cai,Pan Ji,Qingan Yan,Xiangyu Xu,Yi Xu
Synthetic datasets are often used to pretrain end-to-end optical flow networks, due to the lack of a large amount of labeled, real scene data. But major drops in accuracy occur when moving from synthetic to real scenes. How do we better transfer the knowledge learned from synthetic to real domains? To this end, we propose CLIP-Flow, a semi-supervised iterative pseudo labeling framework to transfer the pretraining knowledge to the target real domain. We leverage large-scale, unlabeled real data to facilitate transfer learning with the supervision of iteratively updated pseudo ground truth labels, bridging the domain gap between the synthetic and the real. In addition, we propose a contrastive flow loss on reference features and the warped features by pseudo ground truth flows, to further boost the accurate matching and dampen the mismatching due to motion, occlusion, or noisy pseudo labels. We adopt RAFT as the backbone and obtain an F1-all error of 4.11%, i.e., a 19% error reduction from RAFT (5.10%) and ranking 2nd place at submission on KITTI 2015 benchmark. Our framework can also be extended to other models, e.g., CRAFT, reducing the F1-all error from 4.79% to 4.66% on KITTI 2015 benchmark.

****

CAN: A simple, efficient and scalable contrastive masked autoencoder framework f or learning visual representations

Shlok Kumar Mishra,Joshua David Robinson,Huiwen Chang,David Jacobs,Weicheng Kuo,Aaron Sarna,Aaron Maschinot,Dilip Krishnan

We introduce CAN, a simple, efficient and scalable method for self-supervised learning of visual representations. Our framework is a minimal and conceptually clean synthesis of (C) contrastive learning, (A) masked autoencoders, and (N) the noise prediction approach used in diffusion models. The learning mechanisms are \emph{complementary} to one another: contrastive learning shapes the embedding space across a batch of image samples; masked autoencoders focus on reconstruction of the low-frequency spatial correlations in a single image sample; and noise prediction encourages the reconstruction of the high-frequency components of an image. The combined approach results in a robust, scalable and simple-to-implement algorithm. The training process is symmetric, with 50% of patches in \emph{both views} being masked at random, yielding a considerable efficiency improvement over prior contrastive learning methods. Extensive empirical studies on linear evaluation, finetuning, transfer learning, and robustness demonstrate that our approach achieves strong downstream performance. For instance, when pre-training ViT-B encoders on the curated ImageNet dataset, CAN achieves 74.8% top-1 linear probing accuracy, an absolute improvement of 6.8% over MAE and 1.3% over SimCLR with the same architecture and data augmentations. CAN is especially useful for pre-training on larger uncurated datasets such as JFT-300M: the finetuned performance on ImageNet of our ViT-L model is 85.9%, compared to 85.0% for SimCLR, and 85.4% for MAE. For linear probe on ImageNet, CAN achieves 75.4% compared to 71.8% for SimCLR and 64.1% for MAE. The overall FLOPs load is 41% \emph{lower} than SimCLR\footnote{Our code will be released at \url{www.xxx.yyy}.}.

****

On The Inadequacy of Optimizing Alignment and Uniformity in Contrastive Learning of Sentence Representations

Zhijie Nie,Richong Zhang,Yongyi Mao

Contrastive learning is widely used in areas such as visual representation learning (VRL) and sentence representation learning (SRL). Considering the differences between VRL and SRL in terms of negative sample size and evaluation focus, we believe that the solid findings obtained in VRL may not be entirely carried over to SRL. In this work, we consider the suitability of the decoupled form of contrastive loss, i.e., alignment and uniformity, in SRL. We find a performance gap between sentence representations obtained by jointly optimizing alignment and uniformity on the STS task and those obtained using contrastive loss. Further, we find that the joint optimization of alignment and uniformity during training is prone to overfitting, which does not occur on the contrastive loss. Analyzing them based on the variation of the gradient norms, we find that there is a property of ``gradient dissipation'' in contrastive loss and believe that it is the key to preventing overfitting. We simulate similar "gradient dissipation" of contrastive loss on four optimization objectives of two forms, and achieve the same or even better performance than contrastive loss on the STS tasks, confirming our hypothesis.

****

Bidirectional Learning for Offline Model-based Biological Sequence Design

Can Chen,Yingxue Zhang,Xue Liu,Mark Coates

Offline model-based optimization aims to maximize a black-box objective function with a static dataset of designs and their scores. In this paper, we focus on biological sequence design to maximize some sequence score. A recent approach employs bidirectional learning, combining a forward mapping for exploitation and a backward mapping for constraint, and it relies on the neural tangent kernel (NTK) of an infinitely wide network to build a proxy model. Though effective, the NTK cannot learn features because of its parametrization, and its use prevents the incorporation of powerful pre-trained Language Models (LMs) that can capture the rich biophysical information in millions of biological sequences. We adopt an alternative proxy model, adding a linear head to a pre-trained LM, and propose a linearization scheme. This yields a closed-form loss and also takes into accoun t the biophysical information in the pre-trained LM. In addition, the forward mapping and the backward mapping play different roles and thus deserve different weights during sequence optimization. To achieve this, we train an auxiliary model and leverage its weak supervision signal via a bi-level optimization framework to effectively learn how to balance the two mappings. Further, by extending the framework, we develop the first learning rate adaptation module Adaptive-η, which is compatible with all gradient-based algorithms for offline model-based optimization. Experimental results on DNA/protein sequence design tasks verify the effectiveness of our algorithm. Our code is available https://anonymous.4open.science/r/BIB-ICLR2023-Submission/README.md.

****

Neural Collapse Inspired Feature-Classifier Alignment for Few-Shot Class-Incremental Learning

Yibo Yang,Haobo Yuan,Xiangtai Li,Zhouchen Lin,Philip Torr,Dacheng Tao

Few-shot class-incremental learning (FSCIL) has been a challenging problem as only a few training samples are accessible for each novel class in the new sessions. Finetuning the backbone or adjusting the classifier prototypes trained in the prior sessions would inevitably cause a misalignment between the feature and classifier of old classes, which explains the well-known catastrophic forgetting problem. In this paper, we deal with this misalignment dilemma in FSCIL inspired by the recently discovered phenomenon named neural collapse, which reveals that the last-layer features of the same class will collapse into a vertex, and the vertices of all classes are aligned with the classifier prototypes, which are formed as a simplex equiangular tight frame (ETF). It corresponds to an optimal geometric structure for classification due to the maximized Fisher Discriminant Ratio. We propose a neural collapse inspired framework for FSCIL. A group of classifier prototypes are pre-assigned as a simplex ETF for the whole label space, including the base session and all the incremental sessions. During training, the classifier prototypes are not learnable, and we adopt a novel loss function that drives the features into their corresponding prototypes. Theoretical analysis shows that our method holds the neural collapse optimality and does not break the feature-classifier alignment in an incremental fashion. Experiments on the miniImageNet, CUB-200, and CIFAR-100 datasets demonstrate that our proposed framework outperforms the state-of-the-art performances. Code address: https://github.com/NeuralCollapseApplications/FSCIL

****

Self-conditioned Embedding Diffusion for Text Generation

Robin Strudel,Corentin Tallec,Florent Altché,Yilun Du,Yaroslav Ganin,Arthur Mensch,Will Sussman Grathwohl,Nikolay Savinov,Sander Dieleman,Laurent Sifre,Rémi Leblond

Can continuous diffusion models bring the same performance breakthrough on natural language they did for image generation? To circumvent the discrete nature of text data, we can simply project tokens in a continuous space of embeddings, as is standard in language modeling. We propose Self-conditioned Embedding Diffusion (SED), a continuous diffusion mechanism that operates on token embeddings and allows to learn flexible and scalable diffusion models for both conditional and unconditional text generation. Through qualitative and quantitative evaluation, we show that our text diffusion models generate samples comparable with those produced by standard autoregressive language models — while being in theory more efficient on accelerator hardware at inference time. Our work paves the way for scaling up diffusion models for text, similarly to autoregressive models, and for improving performance with recent refinements to continuous diffusion.

****

Decoupling Concept Bottleneck Model

Rui Zhang,Xingbo Du,Junchi Yan,Shihua Zhang

Concept Bottleneck Model (CBM) is a kind of powerful interpretable neural network, which utilizes high-level concepts to explain model decisions and interact with humans. However, CBM cannot always work as expected due to the troublesome collection and commonplace insufficiency of high-level concepts in real-world scenarios. In this paper, we theoretically reveal that insufficient concept informat ion will induce the mixture of explicit and implicit information, which further leads to the inherent dilemma of concept and label distortions in CBM. Motivated by the proposed theorem, we present Decoupling Concept Bottleneck Model (DCBM), a novel concept-based model decoupling heterogeneous information into explicit and implicit concepts, while still retaining high prediction performance and interpretability. Extensive experiments expose the success in the alleviation of concept/label distortions, where DCBM achieves state-of-the-art performances in both concept and label learning tasks. Especially for situations where concepts are insufficient, DCBM significantly outperforms other models based on concept bottleneck and respectively achieves error rates 24.95% and 20.09% lower than other CBMs on concept/label prediction. Moreover, to express effective human-machine interactions for DCBM, we devise two algorithms based on mutual information (MI) estimation, including forward intervention and backward rectification, which can automatically correct labels and trace back to wrong concepts. The construction of the interaction regime can be formulated as a light min-max optimization problem achieved within minutes. Multiple experiments show that such interactions can effectively promote concept/label accuracy.

****

AQUILA: Communication Efficient Federated Learning with Adaptive Quantization of Lazily-Aggregated Gradients

Zihao Zhao,Yuzhu Mao,Zhenpeng Shi,Muhammad Zeeshan,Yang Liu,Tian Lan,Wenbo Ding

The development and deployment of federated learning (FL) have been bottlenecked by the heavy communication overheads of high-dimensional models between the distributed device nodes and the central server. To achieve better error-communication trade-offs, recent efforts have been made to either adaptively reduce the communication frequency by skipping unimportant updates, a.k.a. lazily-aggregated quantization (LAQ), or adjust the quantization bits for each communication. In this paper, we propose a unifying communication efficient framework for FL based on adaptive quantization of lazily-aggregated gradients (AQUILA), which adaptively adjusts two mutually-dependent factors, the communication frequency, and the quantization level, in a synergistic way. Specifically, we start from a careful investigation of the classical LAQ scheme and formulate AQUILA as an optimization problem where the optimal quantization level per communication is selected by minimizing the model deviation caused by update skipping. Meanwhile, we create a new lazy aggregation strategy to fit the novel quantization criterion better and thus keep the communication frequency at an appropriate level. The effectiveness and convergence of the proposed AQUILA framework are theoretically verified. The experimental results demonstrate that AQUILA can reduce around 60% of overall transmitted bits compared to existing methods while achieving the same level of model accuracy in a number of non-homogeneous FL scenarios, including Non-IID data distribution and heterogeneous model architecture. The proposed AQUILA is highly adaptive and compatible with existing FL settings.

****

Generaling Multimodal Variational Methods to Sets

Jinzhao Zhou,Yiqun Duan,Zhihong Chen,Yu-Cheng Chang,Chin-teng Lin

Making sense of multiple modalities can yield a more comprehensive description of real-world phenomena. However, learning the co-representation of diverse modalities is still a long-standing endeavor in emerging machine learning applications and research. Previous generative approaches for multimodal input approximate a joint-modality posterior by uni-modality posteriors as product-of-experts (PoE) or mixture-of-experts (MoE). We argue that these approximations lead to a defective bound for the optimization process and loss of semantic connection among modalities. This paper presents a novel variational method on sets called the Set Multimodal VAE (SMVAE) for learning a multimodal latent space while handling the missing modality problem. By modeling the joint-modality posterior distribution directly, the proposed SMVAE learns to exchange information between multiple modalities and compensate for the drawbacks caused by factorization. In public datasets of various domains, the experimental results demonstrate that the proposed method is applicable to order-agnostic cross-modal generation while achieving outstanding performance compared to the state-of-the-art multimodal methods. The source code for our method is available online https://anonymous.4open.science/r/SMVAE-9B3C/.

****

Towards a Unified View on Visual Parameter-Efficient Transfer Learning

Bruce Yu,Jianlong Chang,Lingbo Liu,Qi Tian,Chang Wen Chen

Since the release of various large-scale natural language processing (NLP) pre-trained models, parameter efficient transfer learning (PETL) has become a popular paradigm capable of achieving impressive performance on various downstream tasks. PETL aims at making good use of the representation knowledge in the pre-trained large models by fine-tuning a small number of parameters. Recently, it has also attracted increasing attention to developing various PETL techniques for vision tasks. Popular PETL techniques such as Prompt Tuning and Adapter have been proposed for high-level visual downstream tasks such as image classification and video recognition. However, Prefix-tuning remains under-explored for vision tasks. In this work, we intend to adapt large video-based models to downstream tasks with a good parameter-accuracy trade-off. Towards this goal, we propose a framework with a unified view of PETL called visual-PETL (V-PETL) to investigate the effects of different PETL techniques, data scales of downstream domains, positions of trainable parameters, and other aspects affecting the trade-off.

Specifically, we analyze the positional importance of trainable parameters and the differences between NLP and vision tasks in terms of data structures and pre-training mechanisms while implementing various PETL techniques, especially for the under-explored prefix-tuning technique. Based on a comprehensive understanding of the differences between NLP and video data, we propose a new variation of prefix-tuning module called parallel attention (PATT) for video-based downstream tasks.

An extensive empirical analysis on two video datasets via different frozen backbones has been carried and the findings show that the proposed PATT can effectively contribute to other PETL techniques. An effective scheme Swin-BAPAT derived from the proposed V-PETL framework achieves significantly better performance than the state-of-the-art AdaptFormer-Swin with slightly more parameters and outperforms full-tuning with far less parameters.

****

Everyone's Preference Changes Differently: Weighted Multi-Interest Retrieval Model

Hui Shi,Yupeng Gu,Yitong Zhou,Bo Zhao,Sicun Gao,Jishen Zhao

User embeddings (vectorized representations of a user) are essential in recommendation systems. Numerous approaches have been proposed to construct a representation for the user in order to find similar items for retrieval tasks, and they have been proven effective in industrial recommendation systems. Recently people have discovered the power of using multiple embeddings to represent a user, with the hope that each embedding represents the user's interest in a certain topic. With multi-interest representation, it's important to model the user's preference over the different topics and how the preference change with time. However, existing approaches either fail to estimate the user's affinity to each interest or unreasonably assume every interest of every user fades with an equal rate with time, thus hurting the performance of candidate retrieval. In this paper, we propose the Multi-Interest Preference (MIP) model, an approach that not only produces multi-interest for users by using the user's sequential engagement more effectively but also automatically learns a set of weights to represent the preference over each embedding so that the candidates can be retrieved from each interest proportionally. Extensive experiments have been done on various industrial-scale datasets to demonstrate the effectiveness of our approach.

****

Variational Autoencoders with Decremental Information Bottleneck for Disentanglement

Jiantao Wu,Shentong Mo,Muhammad Awais,Sara Atito,xingshen zhang,Lin Wang,Xiang Yang

One major challenge of disentanglement learning with variational autoencoders is

the trade-off between disentanglement and reconstruction fidelity. Previous methods, spreading the conflict of disentanglement and reconstruction in time, will lose the constraint of disentanglement when expanding the information bottleneck, which causes the information diffusion problem. To tackle this issue, we present a novel decremental variational autoencoder with disentanglement-invariant transformations to spread the conflict on multiple latent spaces, termed DeVAE, for balancing disentanglement and reconstruction fidelity by decreasing the information bottleneck of diverse latent spaces gradually. Benefiting from the multiple latent spaces, DeVAE allows simultaneous optimization of multiple objectives to optimize reconstruction while keeping the constraint of disentanglement, avoiding information diffusion. DeVAE is also compatible with large models with high-dimension latent space. Experimental results on dSprites and Shapes3D that DeVAE achieves the best performance on both disentanglement and reconstruction.
****

Volumetric Optimal Transportation by Fast Fourier Transform
Na Lei,DONGSHENG An,Min Zhang,Xiaoyin Xu,David Gu
The optimal transportation map finds the most economical way to transport one probability measure to another, and it has been applied in a broad range of applications in machine learning and computer vision. By the Brenier theory, computing the optimal transport map is equivalent to solving a Monge-Amp\`ere equation, which is highly non-linear. Therefore, the computation of optimal transportation maps is intrinsically challenging.

In this work, we propose a novel and powerful method, the FFT-OT (fast Fourier transform-optimal transport), to compute the 3-dimensional OT problems. The method is based on several key ideas: first, the Monge-Amp\`ere equation is linearized to a sequence of linear elliptic PDEs with spacial and temporal variant coefficients; second, the obliqueness property of optimal transportation maps is reformulated as a Neumann boundary condition; and third, the variant coefficient elliptic PDEs are approximated by constant coefficient elliptic PDEs and solved by FFT on GPUs. We also prove that the algorithm converges linearly, namely the approximation error decreases exponentially fast. Experimental results show that the FFT-OT algorithm is more than a hundred times faster than the conventional methods based on the convex geometry. Furthermore, the method can be directly applied for sampling from complex 3D density functions in machine learning and magnifying the volumetric data in medical imaging.
****

GFlowNets and variational inference
Nikolay Malkin,Salem Lahlou,Tristan Deleu,Xu Ji,Edward J Hu,Katie E Everett,Dinghuai Zhang,Yoshua Bengio
This paper builds bridges between two families of probabilistic algorithms: (hierarchical) variational inference (VI), which is typically used to model distributions over continuous spaces, and generative flow networks (GFlowNets), which have been used for distributions over discrete structures such as graphs. We demonstrate that, in certain cases, VI algorithms are equivalent to special cases of GFlowNets in the sense of equality of expected gradients of their learning objectives. We then point out the differences between the two families and show how these differences emerge experimentally. Notably, GFlowNets, which borrow ideas from reinforcement learning, are more amenable than VI to off-policy training without the cost of high gradient variance induced by importance sampling. We argue that this property of GFlowNets can provide advantages for capturing diversity in multimodal target distributions. Code: https://github.com/GFNOrg/GFN_vs_HVI.
****

Neural Networks and the Chomsky Hierarchy
Gregoire Deletang,Anian Ruoss,Jordi Grau-Moya,Tim Genewein,Li Kevin Wenliang,Elliot Catt,Chris Cundy,Marcus Hutter,Shane Legg,Joel Veness,Pedro A Ortega
Reliable generalization lies at the heart of safe ML and AI. However, understanding when and how neural networks generalize remains one of the most important unsolved problems in the field. In this work, we conduct an extensive empirical study (20'910 models, 15 tasks) to investigate whether insights from the theory of

computation can predict the limits of neural network generalization in practice. We demonstrate that grouping tasks according to the Chomsky hierarchy allows us to forecast whether certain architectures will be able to generalize to out-of-distribution inputs. This includes negative results where even extensive amounts of data and training time never lead to any non-trivial generalization, despite models having sufficient capacity to fit the training data perfectly. Our results show that, for our subset of tasks, RNNs and Transformers fail to generalize on non-regular tasks, LSTMs can solve regular and counter-language tasks, and only networks augmented with structured memory (such as a stack or memory tape) can successfully generalize on context-free and context-sensitive tasks.

****

Neural ePDOs: Spatially Adaptive Equivariant Partial Differential Operator Based Networks

Lingshen He,Yuxuan Chen,Zhengyang Shen,Yibo Yang,Zhouchen Lin

Endowing deep learning models with symmetry priors can lead to a considerable performance improvement. As an interesting bridge between physics and deep learning, the equivariant partial differential operators (PDOs) have drawn much researchers' attention recently. However, to ensure the PDOs translation equivariance, previous works have to require coefficient matrices to be constant and spatially shared for their linearity, which could lead to the sub-optimal feature learning at each position. In this work, we propose a novel nonlinear PDOs scheme that is both spatially adaptive and translation equivariant. The coefficient matrices are obtained by local features through a generator rather than spatially shared. Besides, we establish a new theory on incorporating more equivariance like rotations for such PDOs. Based on our theoretical results, we efficiently implement the generator with an equivariant multilayer perceptron (EMLP). As such equivariant PDOs are generated by neural networks, we call them Neural ePDOs. In experiments, we show that our method can significantly improve previous works with smaller model size in various datasets. Especially, we achieve the state-of-the-art performance on the MNIST-rot dataset with only half parameters of the previous best model.

****

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal,Yuval Alaluf,Yuval Atzmon,Or Patashnik,Amit Haim Bermano,Gal Chechik,Daniel Cohen-or

Text-to-image models offer unprecedented freedom to guide creation through natural language. Yet, it is unclear how such freedom can be exercised to generate images of specific unique concepts, modify their appearance, or compose them in new roles and novel scenes.

In other words, we ask: how can we use language-guided models to turn *our* cat into a painting, or imagine a new product based on *our* favorite toy?

Here we present a simple approach that allows such creative freedom. Using only 3-5 images of a user-provided concept, like an object or a style, we learn to represent it through new ``words" in the embedding space of a frozen text-to-image model.

These ``words" can be composed into natural language sentences, guiding *personalized* creation in an intuitive way.

Notably, we find evidence that a *single* word embedding is sufficient for capturing unique and varied concepts.

We compare our approach to a wide range of baselines, and demonstrate that it can more faithfully portray the concepts across a range of applications and tasks. Our code, data and new words will be available.

****

Cutting Long Gradient Flows: Decoupling End-to-End Backpropagation Based on Supervised Contrastive Learning

Cheng-Kai Wang,Hung-Hsuan Chen

End-to-end backpropagation (BP) is the foundation of current deep learning technology. Unfortunately, as a network becomes deeper, BP becomes inefficient for various reasons. This paper proposes a new methodology for decoupling BP to transf orm a long gradient flow into multiple short ones in order to address the optimization issues caused by long gradient flows. We report thorough experiments conducted to illustrate the effectiveness of our model compared with BP and associated learning (AL), a state-of-the-art methodology for backpropagation decoupling. We will release the source code for the experiments after acceptance.

****

Hierarchical Relational Learning for Few-Shot Knowledge Graph Completion
Han Wu,Jie Yin,Bala Rajaratnam,Jianyuan Guo

Knowledge graphs (KGs) are powerful in terms of their inference abilities, but are also notorious for their incompleteness and long-tail distribution of relations. To address these challenges and expand the coverage of KGs, few-shot KG completion aims to make predictions for triplets involving novel relations when only a few training triplets are provided as reference. Previous methods have focused on designing local neighbor aggregators to learn entity-level information and/or imposing sequential dependency assumption at the triplet level to learn meta relation information. However, pairwise triplet-level interactions and context-level relational information have been largely overlooked for learning meta representations of few-shot relations. In this paper, we propose a hierarchical relational learning method (HiRe) for few-shot KG completion. By jointly capturing three levels of relational information (entity-level, triplet-level and context-level), HiRe can effectively learn and refine the meta representation of few-shot relations, and consequently generalize well to new unseen relations. Extensive experiments on two benchmark datasets validate the superiority of HiRe over state-of-the-art methods. The code of HiRe can be found in supplementary material and will be released after acceptance.

****

Learn to Know Unknowns: A Bionic Memory Network for Unsupervised Anomaly Detection
Jiahao Li,Yiqiang Chen,Yunbing Xing

Is generalization always beneficial? Over-strong generalization induces the model insensitive to anomalies. Unsupervised anomaly detection requires only unlabeled non-anomalous data to learn and generalize normal patterns, which results in a modest reconstruction error when reconstructing normal instances and a significant reconstruction error when reconstructing anomalies. However, over-strong generalization leads to the indistinguishable reconstruction error of normal instances and anomalies, which means that the model well reconstructs the unknown anomalies, resulting in unnoticeable reconstruction error. Inspired by the cascade structure of the hippocampus and cortex in human brain memory, we proposed a re-representation memory network called Random Forgetting Twin Memory (RFTM) to decompose the latent space and introduce a configurable reintegration mechanism to suppress overgeneralization. RFTM shows striking brain-like memory characteristics, which enables the model to know what it does not know. RFTM has the convenience of a single line of code boosting at the model level without adding any additional extra loss terms at the loss function level. RFTM-based models have achieved state-of-the-art experimental results on different public benchmarks.

****

Function-Consistent Feature Distillation
Dongyang Liu,Meina Kan,Shiguang Shan,Xilin CHEN

Feature distillation makes the student mimic the intermediate features of the teacher. Nearly all existing feature-distillation methods use L2 distance or its slight variants as the distance metric between teacher and student features. However, while L2 distance is isotropic w.r.t. all dimensions, the neural network’s operation on different dimensions is usually anisotropic, i.e., perturbations with the same 2-norm but in different dimensions of intermediate features lead to changes in the final output with largely different magnitude. Considering this, we argue that the similarity between teacher and student features should \textit{not} be measured merely based on their appearance (i.e., L2 distance), but should, more importantly, be measured by their difference in function, namely how later layers of the network will read, decode, and process them. Therefore, we pro pose Function-Consistent Feature Distillation (FCFD), which explicitly optimizes the functional similarity between teacher and student features. The core idea of FCFD is to make teacher and student features not only numerically similar, but more importantly produce similar outputs when fed to the later part of the same network. With FCFD, the student mimics the teacher more faithfully and learns more from the teacher. Extensive experiments on image classification and object detection demonstrate the superiority of FCFD to existing methods. Furthermore, we can combine FCFD with many existing methods to obtain even higher accuracy. Our codes are available at https://github.com/LiuDongyang6/FCFD.

****

Multi-User Reinforcement Learning with Low Rank Rewards

Dheeraj Mysore Nagaraj,Suhas S Kowshik,Praneeth Netrapalli,Naman Agarwal,Prateek Jain

In this work, we consider the problem of collaborative multi-user reinforcement learning. In this setting there are multiple users with the same state-action space and transition probabilities but with different rewards. Under the assumption that the reward matrix of the N users has a low-rank structure -- a standard and practically successful assumption in the offline collaborative filtering setting-- the question is can we design algorithms with significantly lower sample complexity compared to the ones that learn the MDP individually for each user. Our main contribution is an algorithm which explores rewards collaboratively with N user-specific MDPs and can learn rewards efficiently in two key settings: tabular MDPs and linear MDPs. When N is large and the rank is constant, the sample complexity per MDP depends logarithmically over the size of the state-space, which represents an exponential reduction (in the state-space size) when compared to the standard ``non-collaborative'' algorithms.

****

The Devil is in the Wrongly-classified Samples: Towards Unified Open-set Recognition

Jun CEN,Di Luan,Shiwei Zhang,Yixuan Pei,Yingya Zhang,Deli Zhao,Shaojie Shen,Qifeng Chen

Open-set Recognition (OSR) aims to identify test samples whose classes are not seen during the training process. Recently, Unified Open-set Recognition (UOSR) has been proposed to reject not only unknown samples but also known but wrongly classified samples, which tends to be more practical in real-world applications. In this paper, we deeply analyze the UOSR task under different training and evaluation settings to shed light on this promising research direction. For this purpose, we first evaluate the UOSR performance of several OSR methods and show a significant finding that the uncertainty distribution of almost all these methods is actually closer to the expectation of UOSR than OSR. We show that the reason lies in the known but wrongly classified samples, as their uncertainty distribution is extremely close to unknown samples rather than known and correctly classified samples. Second, we analyze how the two training settings of OSR (i.e., pre-training and outlier exposure) influence the UOSR. We find although they are both beneficial for distinguishing known and correctly classified samples from unknown samples, pre-training is also helpful for identifying known but wrongly classified samples while outlier exposure is not. In addition to different training settings, we also formulate a new evaluation setting for UOSR which is called few-shot UOSR, where only one or five samples per unknown class are available during evaluation to help identify unknown samples. We propose FS-KNNS for the few-shot UOSR to achieve state-of-the-art performance under all settings.

****

Approximated Anomalous Diffusion: Gaussian Mixture Score-based Generative Models

Hengyuan Ma,Li Zhang,Xiatian Zhu,Jianfeng Feng

Score-based generative models (SGMs) can generate high-quality samples via Langevin dynamics with a drift term and a diffusion term (Gaussian noise) iteratively calculated and added to a sample until convergence. In biological systems, it is observed that the neural population can conduct heavy-tailed L\'{e}vy dynamics for sampling-based probabilistic representation through neural fluctuations. Critically, unlike the existing sampling process of SGMs, L\'{e}vy dynamics can pr oduce both large jumps and small roaming to explore the sampling space, resulting in better sampling results than Langevin dynamics with a lacking of large jumps. Motivated by this contrast, we explore a new class of SGMs with the sampling based on the L\'{e}vy dynamics. However, exact numerical simulation of the L\'{e}vy dynamics is significantly more challenging and intractable. We hence propose an approximation solution by leveraging Gaussian mixture noises during training to achieve the desired large jumps and small roaming properties. Theoretically, GM-SGMs conduct a probabilistic graphical model used by empirical Bayes for sampling, expanding the maximum a posteriori (MAP) estimation applied by conventional SGMs. Expensive experiments on the challenging image generation tasks show that our GM-SGMs exhibit superior sampling quality over prior art SGMs across various sampling iterations.

****

MCAL: Minimum Cost Human-Machine Active Labeling

Hang Qiu,Krishna Chintalapudi,Ramesh Govindan

Today, ground-truth generation uses data sets annotated by cloud-based annotation services. These services rely on human annotation, which can be prohibitively expensive. In this paper, we consider the problem of hybrid human-machine labeling, which trains a classifier to accurately auto-label part of the data set. However, training the classifier can be expensive too. We propose an iterative approach that minimizes total overall cost by, at each step, jointly determining which samples to label using humans and which to label using the trained classifier. We validate our approach on well known public data sets such as Fashion-MNIST, CIFAR-10, CIFAR-100, and ImageNet. In some cases, our approach has 6× lower overall cost relative to human labeling the entire data set, and is always cheaper than the cheapest competing strategy.

****

SegNeRF: 3D Part Segmentation with Neural Radiance Fields

Jesus Zarzar,Sara Rojas Martinez,Silvio Giancola,Bernard Ghanem

Recent advances in Neural Radiance Fields (NeRF) boast impressive performances for generative tasks such as novel view synthesis and 3D reconstruction. Methods based on neural radiance fields are able to represent the 3D world implicitly by relying exclusively on posed images. Yet, they have seldom been explored in the realm of discriminative tasks such as 3D part segmentation. In this work, we attempt to bridge that gap by proposing SegNeRF: a neural field representation that integrates a semantic field along with the usual radiance field. SegNeRF inherits from previous works the ability to perform novel view synthesis and 3D reconstruction, and enables 3D part segmentation from a few images. Our extensive experiments on PartNet show that SegNeRF is capable of simultaneously predicting geometry, appearance, and semantic information from posed images, even for unseen objects. The predicted semantic fields allow SegNeRF to achieve an average mIoU of 30.30% for 2D novel view segmentation, and 37.46% for 3D part segmentation, boasting competitive performance against point-based methods by using only a few posed images. Additionally, SegNeRF is able to generate an explicit 3D model from a single image of an object taken in the wild, with its corresponding part segmentation.

****

EyeDAS: Securing Perception of Autonomous Cars Against the Stereoblindness Syndrome

Efrat Levy,Ben Nassi,Raz Swissa,Yuval Elovici

The ability to detect whether an object is a 2D or 3D object is extremely important in autonomous driving, since a detection error can have life-threatening consequences, endangering the safety of the driver, passengers, pedestrians, and others on the road.

Methods proposed to distinguish between 2 and 3D objects (e.g., liveness detection methods) are not suitable for autonomous driving, because they are object dependent or do not consider the constraints associated with autonomous driving (e.g., the need for real-time decision-making while the vehicle is moving).

In this paper, we present EyeDAS, a novel few-shot learning-based method aimed at securing an object detector (OD) against the threat posed by the stereoblindne ss syndrome (i.e., the inability to distinguish between 2D and 3D objects).

We evaluate EyeDAS's real-time performance using 2,000 objects extracted from seven YouTube video recordings of street views taken by a dash cam from the driver's seat perspective.

When applying EyeDAS to seven state-of-the-art ODs as a countermeasure, EyeDAS was able to reduce the 2D misclassification rate from 71.42-100% to 2.4% with a 3D misclassification rate of 0% (TPR of 1.0).

We also show that EyeDAS outperforms the baseline method and achieves an AUC of over 0.999.

****

Learnable Topological Features For Phylogenetic Inference via Graph Neural Networks

Cheng Zhang

Structural information of phylogenetic tree topologies plays an important role in phylogenetic inference. However, finding appropriate topological structures for specific phylogenetic inference tasks often requires significant design effort and domain expertise. In this paper, we propose a novel structural representation method for phylogenetic inference based on learnable topological features. By combining the raw node features that minimize the Dirichlet energy with modern graph representation learning techniques, our learnable topological features can provide efficient structural information of phylogenetic trees that automatically adapts to different downstream tasks without requiring domain expertise. We demonstrate the effectiveness and efficiency of our method on a simulated data tree probability estimation task and a benchmark of challenging real data variational Bayesian phylogenetic inference problems.

****

Double dynamic sparse training for GANs

Yite Wang,Jing Wu,Naira Hovakimyan,Ruoyu Sun

The past decade has witnessed a drastic increase in modern deep neural networks (DNNs) size, especially for generative adversarial networks (GANs). Since GANs usually suffer from high computational complexity, researchers have shown an increased interest in applying pruning methods to reduce the training and inference costs of GANs. Among different pruning methods invented for supervised learning, dynamic sparse training (DST) has gained increasing attention recently as it enjoys excellent training efficiency with comparable performance to post-hoc pruning. Hence, applying DST on GANs, where we train a sparse GAN with a fixed parameter count throughout training, seems to be a good candidate for reducing GAN training costs. However, a few challenges, including the degrading training instability, emerge due to the adversarial nature of GANs. Hence, we introduce a quantity called balance ratio (BR) to quantify the balance of the generator and the discriminator. We conduct a series of experiments to show the importance of BR in understanding sparse GAN training. Building upon single dynamic sparse training (SDST), where only the generator is adjusted during training, we propose double dynamic sparse training (DDST) to control the BR during GAN training. Empirically, DDST automatically determines the density of the discriminator and greatly boosts the performance of sparse GANs on multiple datasets.

****

Hardware-restriction-aware training (HRAT) for memristor neural networks

Zhimin Tang,Rujie Zhao,Linkai Luo,Haibo Wang,Chao Lu

Memristor neural network (MNN), which utilizes memristor crossbars for vector-matrix multiplication, has huge advantages in terms of scalability and energy efficiency for neuromorphic computing. MNN weights are usually trained offline and then deployed as memristor conductances through a sequence of programming voltage pulses. Although weight uncertainties caused by process variation have been addressed in variation-aware training algorithms, efficient design and training of MNNs have not been systematically explored to date. In this work, we propose Hardware-Restriction-Aware Training (HRAT), which takes into account various non-negligible limitations and non-idealities of memristor devices, circuits, and systems. HRAT considers MNN's realistic behavior and circuit restrictions during offline training, thereby bridging the gap between offline training and hardware de

ployment. HRAT uses a new batch normalization (BN) fusing strategy to align the distortion caused by hardware restrictions between offline training and hardware inference. This not only improves inference accuracy but also eliminates the need for dedicated circuitry for BN operations. Furthermore, most normal scale signals are limited in amplitude due to the restriction of non-destructive threshold voltage of memristors. To avoid input signal distortion of memristor crossbars, HRAT dynamically adjusts the input signal magnitude during training using a learned scale factor. These scale factors can be incorporated into the parameters of linear operation together with fused BN, so no additional signal scaling circuits are required. To evaluate the proposed HRAT methodology, FC-4 and LeNet-5 on MNIST are firstly trained by HRAT and then deployed in hardware. Hardware simulations match well with the offline HRAT results. We also carried out various experiments using VGG-16 on the CIFAR datasets. The study shows that HRAT leads to high-performance MNNs without device calibration or on-chip training, thus greatly facilitating commercial MNN deployment.
****

DifFace: Blind Face Restoration with Diffused Error Contraction

Zongsheng Yue,Chen Change Loy

While deep learning-based methods for blind face restoration have achieved unprecedented success, they still suffer from two major limitations. First, most of them deteriorate when facing complex degradations out of their training data. Second, these methods require multiple constraints, e.g., fidelity, perceptual, and adversarial losses, which requires laborious hyper-parameters tuning to stabilize and balance their influences. In this work, we propose a novel method named DifFace, being able to cope with unseen and complex degradations more gracefully without complicated loss designs. The key of our method is to establish a posterior distribution from the observed low-quality (LQ) image to its high-quality (HQ) counterpart. In particular, we design a transition distribution from the LQ image to the intermediate state of a pre-trained diffusion model and then gradually transmit from this intermediate state to the HQ target by recursively applying a pre-trained diffusion model. The transition distribution only relies on a restoration backbone that is trained with L2 loss on some synthetic data, which favorably avoids the cumbersome training process in existing methods. Moreover, the transition distribution is capable of contracting the error of the restoration backbone and thus makes our method more robust to unknown degradations. Comprehensive experiments show that DifFace is superior to current state-of-the-art methods, especially in cases with severe degradations. Code and model will be released.
****

Fairness-aware Contrastive Learning with Partially Annotated Sensitive Attributes

Fengda Zhang,Kun Kuang,Long Chen,Yuxuan Liu,Chao Wu,Jun Xiao

Learning high-quality representation is important and essential for visual recognition. Unfortunately, traditional representation learning suffers from fairness issues since the model may learn information of sensitive attributes. Recently, a series of studies have been proposed to improve fairness by explicitly decorrelating target labels and sensitive attributes. Most of these methods, however, rely on the assumption that fully annotated labels on target variable and sensitive attributes are available, which is unrealistic due to the expensive annotation cost. In this paper, we investigate a novel and practical problem of Fair Unsupervised Representation Learning with Partially annotated Sensitive labels (FURL-PS). FURL-PS has two key challenges: 1) how to make full use of the samples that are not annotated with sensitive attributes; 2) how to eliminate bias in the dataset without target labels. To address these challenges, we propose a general Fairness-aware Contrastive Learning (FairCL) framework consisting of two stages. Firstly, we generate contrastive sample pairs, which share the same visual information apart from sensitive attributes, for each instance in the original dataset. In this way, we construct a balanced and unbiased dataset. Then, we execute fair contrastive learning by closing the distance between representations of contrastive sample pairs. Besides, we also propose an unsupervised way to balance

the utility and fairness of learned representations by feature reweighting. Extensive experimental results illustrate the effectiveness of our method in terms of fairness and utility, even with very limited sensitive attributes and serious data bias.
****
Training Instability and Disharmony Between ReLU and Batch Normalization
Inyoung Paik,Jaesik Choi
Deep neural networks based on batch normalization and ReLU-like activation functions experience instability during early stages of training owing to the high gradient induced by temporal gradient explosion. ReLU reduces the variance by more than the expected amount and batch normalization amplifies the gradient during its recovery. In this paper, we explain the explosion of a gradient mathematically while the forward propagation remains stable, and also the alleviation of the problem during training. Based on this, we propose a Layer-wise Asymmetric Learning rate Clipping (LALC) algorithm, which outperforms existing learning rate scaling methods in large batch training and can also be used to replace WarmUp in small batch training.
****
Rotamer Density Estimator is an Unsupervised Learner of the Effect of Mutations on Protein-Protein Interaction
Shitong Luo,Yufeng Su,Zuofan Wu,Chenpeng Su,Jian Peng,Jianzhu Ma
Protein-protein interactions are crucial to many biological processes, and predicting the effect of amino acid mutations on binding is important for protein engineering. While data-driven approaches using deep learning have shown promise, the scarcity of annotated experimental data remains a major challenge. In this work, we propose a new approach that predicts mutational effects on binding using the change in conformational flexibility of the protein-protein interface. Our approach, named Rotamer Density Estimator (RDE), employs a flow-based generative model to estimate the probability distribution of protein side-chain conformations and uses entropy to measure flexibility. RDE is trained solely on protein structures and does not require the supervision of experimental values of changes in binding affinities. Furthermore, the unsupervised representations extracted by RDE can be used for downstream neural network predictions with even greater accuracy. Our method outperforms empirical energy functions and other machine learning-based approaches.
****
Dilated convolution with learnable spacings
Ismail Khalfaoui Hassani,Thomas Pellegrini,Timothée Masquelier
Recent works indicate that convolutional neural networks (CNN) need large receptive fields (RF) to compete with visual transformers and their attention mechanism. In CNNs, RFs can simply be enlarged by increasing the convolution kernel sizes. Yet the number of trainable parameters, which scales quadratically with the kernel's size in the 2D case, rapidly becomes prohibitive, and the training is notoriously difficult. This paper presents a new method to increase the RF size without increasing the number of parameters. The dilated convolution (DC) has already been proposed for the same purpose. DC can be seen as a convolution with a kernel that contains only a few non-zero elements placed on a regular grid. Here we present a new version of the DC in which the spacings between the non-zero elements, or equivalently their positions, are no longer fixed but learnable via backpropagation thanks to an interpolation technique. We call this method “Dilated Convolution with Learnable Spacings” (DCLS) and generalize it to the n-dimensional convolution case. However, our main focus here will be on the 2D case. We first tried our approach on ResNet50: we drop-in replaced the standard convolutions with DCLS ones, which increased the accuracy of ImageNet1k classification at iso-parameters, but at the expense of the throughput. Next, we used the recent ConvNeXt state-of-the-art convolutional architecture and drop-in replaced the depthwise convolutions with DCLS ones. This not only increased the accuracy of ImageNet1k classification but also of typical downstream and robustness tasks, again at iso-parameters but this time with negligible cost on throughput, as ConvNeXt uses separable convolutions. Conversely, classic DC led to poor performance wit

h both ResNet50 and ConvNeXt. The code of the method is based on PyTorch and available at: https://github.com/K-H-Ismail/Dilated-Convolution-with-Learnable-Spacings-PyTorch.

****

PatchDCT: Patch Refinement for High Quality Instance Segmentation

Qinrou Wen,Jirui Yang,Xue Yang,Kewei Liang

High-quality instance segmentation has shown emerging importance in computer vision. Without any refinement, DCT-Mask directly generates high-resolution masks by compressed vectors. To further refine masks obtained by compressed vectors, we propose for the first time a compressed vector based multi-stage refinement framework. However, the vanilla combination does not bring significant gains, because changes in some elements of the DCT vector will affect the prediction of the entire mask. Thus, we propose a simple and novel method named PatchDCT, which separates the mask decoded from a DCT vector into several patches and refines each patch by the designed classifier and regressor. Specifically, the classifier is used to distinguish mixed patches from all patches, and to correct previously mispredicted foreground and background patches. In contrast, the regressor is used for DCT vector prediction of mixed patches, further refining the segmentation quality at boundary locations. Experiments on COCO show that our method achieves 2.0\%, 3.2\%, 4.5\% AP and 3.4\%, 5.3\%, 7.0\% Boundary AP improvements over Mask-RCNN on COCO, LVIS, and Cityscapes, respectively. It also surpasses DCT-Mask by 0.7\%, 1.1\%, 1.3\% AP and 0.9\%, 1.7\%, 4.2\% Boundary AP on COCO, LVIS and Cityscapes. Besides, the performance of PatchDCT is also competitive with other state-of-the-art methods.

****

Global Prototype Encoding for Incremental Video Highlights Detection

Sen Pei,Shixiong Xu,Ye Yuan,Jiashi Feng,Xiaohui Shen,Xiaojie Jin

Video highlights detection (VHD) is an active research field in computer vision, aiming to locate the most user-appealing clips given raw video inputs. However, most VHD methods are based on the closed world assumption, \emph{i.e.}, a fixed number of highlight categories is defined in advance and all training data are available beforehand. Consequently, existing methods have poor scalability with respect to increasing highlight domains and training data. To address above issues, we propose a novel video highlight detection method named \textbf{G}lobal \textbf{P}rototype \textbf{E}ncoding (GPE) to learn incrementally for adapting to new domains via parameterized prototypes. To facilitate this new research direction, we collect a finely annotated dataset termed \emph{LiveFood}, including over 5,100 live gourmet videos that consist of four domains: \emph{cooking}, \emph{eating}, \emph{ingredients} and \emph{presentation}. To the best of our knowledge, this is the first work to explore video highlight detection in the incremental learning setting, opening up new land to apply VHD for practical scenarios where both the concerned highlight domains and training data increase over time. We demonstrate the effectiveness of GPE through extensive experiments. Notably, GPE surpasses popular domain-incremental learning methods on \emph{LiveFood}, achieving significant mAP improvements on all domains. The code and dataset will be made publicly available.

****

WaGI: Wavelet-based GAN Inversion for Preserving High-Frequency Image Details

SeungJun Moon,Chaewon Kim,Gyeong-Moon Park

Recent GAN inversion models focus on preserving image-specific details through various methods, e.g., generator tuning or feature mixing.

While those are helpful for preserving details compared to naive low-rate latent inversion, they still fail to maintain high-frequency features precisely. In this paper, we point out that existing GAN inversion models have inherent limitations in both structural and training aspects, which preclude the delicate reconstruction of high-frequency features. Especially, we prove that the widely-used loss term in GAN inversion, i.e., is biased to mainly reconstructing low-frequency features. To overcome this problem, we propose a novel GAN inversion model, coined WaGI, which enables handling high-frequency features explicitly, by using a novel wavelet-based loss term and a newly proposed wavelet fusion scheme. To the best of our knowledge, WaGI is the first approach to interpret GAN inversion in the frequency domain. We demonstrate that WaGI shows outstanding results on both inversion and editing, compared to existing state-of-the-art GAN inversion models. Especially, WaGI robustly preserves high-frequency features of images even in the editing scenario. We will release our code with the pre-trained model after the review.

****

Neural-Symbolic Recursive Machine for Systematic Generalization

Qing Li,Yixin Zhu,Yitao Liang,Ying Nian Wu,Song-Chun Zhu,Siyuan Huang

Despite the tremendous success, existing machine learning models still fall short of human-like systematic generalization—learning compositional rules from limited data and applying them to unseen combinations in various domains. We propose Neural-Symbolic Recursive Machine (NSR) to tackle this deficiency. The core representation of NSR is a Grounded Symbol System (GSS) with combina- torial syntax and semantics, which entirely emerges from training data. Akin to the neuroscience studies suggesting separate brain systems for perceptual, syntactic, and semantic processing, NSR implements analogous separate modules of neural perception, syntactic parsing, and semantic reasoning, which are jointly learned by a deduction-abduction algorithm. We prove that NSR is expressive enough to model various sequence-to-sequence tasks. Superior systematic generalization is achieved via the inductive biases of equivariance and recursiveness embedded in NSR. In experiments, NSR achieves state-of-the-art performance in three benchmarks from different domains: SCAN for semantic parsing, PCFG for string manipulation, and HINT for arithmetic reasoning. Specifically, NSR achieves 100% generalization accuracy on SCAN and PCFG and outperforms state-of-the-art models on HINT by about 23%. Our NSR demonstrates stronger generalization than pure neural networks due to its symbolic representation and inductive biases. NSR also demonstrates better transferability than existing neural-symbolic approaches due to less domain-specific knowledge required.

****

ChiroDiff: Modelling chirographic data with Diffusion Models

Ayan Das,Yongxin Yang,Timothy Hospedales,Tao Xiang,Yi-Zhe Song

Generative modelling over continuous-time geometric constructs, a.k.a $chirographic\ data$ such as handwriting, sketches, drawings etc., have been accomplished through autoregressive distributions. Such strictly-ordered discrete factorization however falls short of capturing key properties of chirographic data -- it fails to build holistic understanding of the temporal concept due to one-way visibility (causality). Consequently, temporal data has been modelled as discrete token sequences of fixed sampling rate instead of capturing the true underlying concept. In this paper, we introduce a powerful model-class namely Denoising\ Diffusion\ Probabilistic\ Models or DDPMs for chirographic data that specifically addresses these flaws. Our model named "ChiroDiff", being non-autoregressive, learns to capture holistic concepts and therefore remains resilient to higher temporal sampling rate up to a good extent. Moreover, we show that many important downstream utilities (e.g. conditional sampling, creative mixing) can be flexibly implemented using ChiroDiff. We further show some unique use-cases like stochastic vectorization, de-noising/healing, abstraction are also possible with this model-class. We perform quantitative and qualitative evaluation of our framework on relevant datasets and found it to be better or on par with competing approaches.

****

Active Topological Mapping by Metric-Free Exploration via Task and Motion Imitation

Yuhang He,Irving Fang,Yiming Li,Chen Feng

Topological map is an effective environment representation for visual navigation. It is a graph of image nodes and spatial neighborhood edges without metric information such as global or relative agent poses. However, currently such a map construction relies on either less-efficient random exploration or more demanding training involving metric information. To overcome these issues, we propose active topological mapping (ATM), consisting of an active visual exploration and a topological mapping by visual place recognition. Our main novelty is the simple

and lightweight active exploration policy that works entirely in the image feature space involving no metric information. More specifically, ATM's metric-free exploration is based on task and motion planning (TAMP). The task planner is a recurrent neural network using the latest local image observation sequence to hallucinate a feature as the next-step best exploration goal. The motion planner then fuses the current and the hallucinated feature to generate an action taking the agent towards the hallucinated feature goal. The two planners are jointly trained via deeply-supervised imitation learning from expert exploration demonstrations. Extensive experiments in both exploration and navigation tasks on the photo-realistic Gibson and MP3D datasets validate ATM's effectiveness and generalizability.
****

SoundCount: Sound Counting from Raw Audio with Dyadic Decomposition Neural Network

Yuhang He,Zhuangzhuang Dai,Niki Trigoni,Andrew Markham

In this paper, we study an underexplored, yet important and challenging problem: counting the number of distinct sound events in data characterized by a high degree of polyphonicity and spectral overlap. A key example is counting individual bird calls in bioacoustic data, from which biodiversity can be estimated. We do so by systematically proposing a novel end-to-end trainable neural network, designing new evaluation protocols, quantifying the difficulty of counting depending on sound polyphonicity, and creating a new dataset tailored for concurrent sound event counting. Unlike existing methods that all apply frequency-selective filters on the raw waveform in a one-stage manner, our neural network progressively decomposes the raw waveform dyadically in frequency domain. Taking inspiration from wavelet decomposition, intermediate waveforms convolved by a parent filter are successively processed by a pair of children filters that evenly split the parent filter's carried frequency response. An energy gain normalization module is introduced to normalize received sound events' loudness variance and spectrum overlap. The network is fully convolutional and parameter-frugal so it is light-weight and computationally efficient. We further design a set of polyphony-aware metrics to quantify sound counting difficulty level from different perspectives. To show the efficiency and generalization of our method (we call DyDecNet), we do experiments on both bioacoustic bird sound (both synthetic and real-world sound), telephone-ring sound and music sound data. Comprehensive experiment results show our method outperforms existing sound event detection (SED) methods significantly. The dyadic decomposition front-end network can be used by existing methods to improve their performance accordingly.
****

SoundNeRirF: Receiver-to-Receiver Sound Neural Room Impulse Response Field

Yuhang He,Jia-Xing Zhong,Zhuangzhuang Dai,Niki Trigoni,Andrew Markham

We present SoundNeRirF, a framework that learns a continuous receiver-to-receiver neural room impulse response field~(r2r-RIR) to help robot efficiently predict the sound to be heard at novel locations. It represents a room acoustic scene as a continuous 6D function, whose input is a reference receiver's 3D position and a target receiver's 3D position, and whose outputs are an inverse room impulse response~(inverse-RIR) and a forward room impulse response~(forward-RIR) that jointly project the sound from the reference position to the target position. SoundNeRirF requires knowledge of neither sound source (e.g. location and number of sound sources) nor room acoustic properties~(e.g. room size, geometry, materials). Instead, it merely depends on a sparse set of sound receivers' positions, as well as the recorded sound at each position. We instantiate the continuous 6D function as multi-layer perceptrons~(MLP), so it is fully differentiable and continuous at any spatial position. SoundNeRirF is encouraged, during the training stage, to implicitly encode the interaction between sound sources, receivers and room acoustic properties by minimizing the discrepancy between the predicted sound and the truly heard sound at the target position. During inference, the sound at a novel position is predicted by giving a reference position and the corresponding reference sound. Extensive experiments on both synthetic and real-world datasets show SoundNeRirF is capable of predicting high-fidelity and audio-realis tic sound that fully captures room reverberation characteristics, significantly outperforming existing methods in terms of accuracy and efficiency.

****

Towards Sustainable Self-supervised Learning

Shang-Hua Gao,Pan Zhou,Ming-Ming Cheng,Shuicheng YAN

Though increasingly training-expensive, most self-supervised learning (SSL) models have repeatedly been trained from scratch but not fully utilized since only a few SOTAs are adopted for downstream tasks. In this work, we explore a sustainable SSL framework with two major challenges: i) learning a stronger new SSL model based on the existing pretrained SSL model in a cost-friendly manner, ii) allowing the training of the new model to be compatible with various base models. We propose a Target-Enhanced Conditional (TEC) scheme, which introduces two components to existing mask-reconstruction based SSL. Firstly, we introduce patch-relation enhanced targets to encourage the new model to learn semantic-relation knowledge from the base model using incomplete inputs. This hardening and target-enhancing could help the new model surpass the base model, since they enforce additional patch relation modeling to handle incomplete input. Secondly, we introduce a conditional adapter that adaptively adjusts new model prediction to align with the target of each base model. Experimental results show that our TEC scheme can accelerate the learning speed and also improve SOTA SSL models, e.g., MAE and iBOT, taking an explorative step towards sustainable SSL.

****

Real-Time Image Demoir$\acute{e}$ing on Mobile Devices

Yuxin Zhang,Mingbao Lin,Xunchao Li,Han Liu,Guozhi Wang,Fei Chao,Ren Shuai,Yafei Wen,Xiaoxin Chen,Rongrong Ji

Moir$\acute{e}$ patterns appear frequently when taking photos of digital screens, drastically degrading the image quality. Despite the advance of CNNs in image demoir$\acute{e}$ing, existing networks are with heavy design, causing massive computation burden for mobile devices. In this paper, we launch the first study on accelerating demoir$\acute{e}$ing networks and propose a dynamic demoir$\acute{e}$ing acceleration method (DDA) towards a real-time deployment on mobile devices. Our stimulus stems from a simple-yet-universal fact that moir${\'e}$ patterns often unbalancedly distribute across an image. Consequently, excessive computation is wasted upon non-moir$\acute{e}$ areas. Therefore, we reallocate computation costs in proportion to the complexity of image patches. In order to achieve this aim, we measure the complexity of an image patch by a novel moir$\acute{e}$ prior that considers both colorfulness and frequency information of moir$\acute{e}$ patterns. Then, we restore higher-complex image patches using larger networks and the lower-complex ones are assigned with smaller networks to relieve the computation burden. At last, we train all networks in a parameter-shared supernet paradigm to avoid additional parameter burden. Extensive experiments on several benchmarks demonstrate the efficacy of our DDA. In addition, the acceleration evaluated on the VIVO X80 Pro smartphone equipped with the chip of Snapdragon 8 Gen 1 also shows that our method can drastically reduce the inference time, leading to a real-time image demoir$\acute{e}$ing on mobile devices. Source codes and models are released at https://github.com/zyxxmu/DDA.

****

Domain Generalization via Independent Regularization from Early-branching Networks

Liang Chen,Yong Zhang,Yibing Song,Jue Wang,Lingqiao Liu

Learning domain-invariant feature representations is critical for achieving domain generalization, where a model is required to perform well on unseen domains. The critical challenge is that standard training often results in entangled domain-invariant and domain-specific features. To address this issue, we use a dual-branching network to learn two features, one for the domain classification problem and the other for the original target classification problem, and the feature of the latter is required to be independent of the former. While this idea seems straightforward, we show that several factors need to be carefully considered for it to work effectively. In particular, we investigate different branching st ructures and discover that the common practice of using a shared base feature extractor with two lightweight prediction heads is detrimental to the performance. Instead, a simple early-branching architecture, where the domain classification and target classification branches share the first few blocks while diverging thereafter, leads to better results. Moreover, we also incorporate a random style augmentation scheme as an extension to further unleash the power of the proposed method, which can be seamlessly integrated into the dual-branching network by our loss terms. Such an extension gives rise to an effective domain generalization method. Experimental results show that the proposed method outperforms state-of-the-art domain generalization methods on various benchmark datasets.

****

AutoSKDBERT: Learn to Stochastically Distill BERT

Zixiang Ding,GUO-QING JIANG,Shuai Zhang,Lin Guo,Wei Lin

In this paper, we propose AutoSKDBERT, a new knowledge distillation paradigm for BERT compression, that stochastically samples a teacher from a predefined teacher team following a categorical distribution in each step, to transfer knowledge into student. AutoSKDBERT aims to discover the optimal categorical distribution which plays an important role to achieve high performance. The optimization procedure of AutoSKDBERT can be divided into two phases: 1) phase-1 optimization distinguishes effective teachers from ineffective teachers, and 2) phase-2 optimization further optimizes the sampling weights of the effective teachers to obtain satisfactory categorical distribution. Moreover, after phase-1 optimization completion, AutoSKDBERT adopts teacher selection strategy to discard the ineffective teachers whose sampling weights are assigned to the effective teachers. Particularly, to alleviate the gap between categorical distribution optimization and evaluation, we also propose a stochastic single-weight optimization strategy which only updates the weight of the sampled teacher in each step. Extensive experiments on GLUE benchmark show that the proposed AutoSKDBERT achieves state-of-the-art score compared to previous compression approaches on several downstream tasks, including pushing MRPC F1 and accuracy to 93.2 (0.6 point absolute improvement) and 90.7 (1.2 point absolute improvement), RTE accuracy to 76.9 (2.9 point absolute improvement).

****

Training A Multi-stage Deep Classifier with Feedback Signals

Chao Xu,Yu Yang,Rongzhao Wang,Guan Wang,Bojia Lin

Multi-Stage Classifier (MSC) - several classifiers working sequentially in an arranged order and classification decision is partially made at each step - is widely used in industrial applications for various resource limitation reasons. The classifiers of a multi-stage process are usually Neural Network (NN) models trained independently or in their inference order without considering the signals from the latter stages. Aimed at two-stage binary classification process, the most common type of MSC, we propose a novel training framework, named Feedback Training. The classifiers are trained in an order reverse to their actual working order, and the classifier at the later stage is used to guide the training of initial-stage classifier via a sample weighting method. We experimentally show the efficacy of our proposed approach, and its great superiority under the scenario of few-shot training.

****

Is Self-Supervised Contrastive Learning More Robust Than Supervised Learning?

Yuanyi Zhong,Haoran Tang,Junkun Chen,Jian Peng,Yu-Xiong Wang

Prior work on self-supervised contrastive learning has primarily focused on evaluating the recognition accuracy, but has overlooked other behavioral aspects. In addition to accuracy, distributional robustness plays a critical role in the reliability of machine learning models. We design and conduct a series of robustness tests to quantify the behavioral differences between contrastive learning and supervised learning to downstream and pre-training data distribution changes. These tests leverage data corruptions at multiple levels, ranging from pixel-level distortion to patch-level shuffling and to dataset-level distribution shift, including both natural and unnatural corruptions. Our tests unveil intriguing robustness behaviors of contrastive and supervised learning: while we generally obs erve that contrastive learning is more robust than supervised learning under downstream corruptions, we surprisingly discover the robustness vulnerability of contrastive learning under pixel and patch level corruptions during pre-training. Furthermore, we observe the higher dependence of contrastive learning on spatial image coherence information during pre-training, e.g., it is particularly sensitive to global patch shuffling. We explain these results by connecting to feature space uniformity and data augmentation. Our analysis has implications in improving the downstream robustness of supervised learning, and calls for more studies on understanding contrastive learning.

****

An Empirical Study of Metrics to Measure Representational Harms in Pre-Trained Language Models

Saghar Hosseini,Ahmed Hassan Awadallah,Hamid Palangi

Large-scale Pre-Trained Language Models (PTLMs) capture knowledge from massive human-written data which contains latent societal biases and toxic contents. In this paper, we leverage the primary task of PTLMs, i.e. language modeling, and propose a new metric to quantify manifested implicit representational harms in PTLMs towards 13 marginalized demographics. Using this metric, we conducted an empirical analysis of 24 widely used PTLMs. Our analysis provides insights into the correlation between the proposed metric in this work and other related fairness metrics. We observe that our metric correlates with the majority of gender-specific fairness metrics in the literature. Through extensive experiments, we explore the connections between PTLMs architectures and representational harms across two dimensions: depth and width of the networks. We found that prioritizing depth over width, mitigates representational harms in some PTLMs.

****

Unsupervised Learning of Causal Relationships from Unstructured Data

Marian Longa,Joao F. Henriques

Endowing deep neural networks with the ability to reason about cause and effect would be an important step to make them more robust and interpretable. In this work we propose a variational framework that allows deep networks to learn latent variables and their causal relationships from unstructured data, with no supervision, or labeled interventions. Starting from an abstract Structural Equation Model (SEM), we show that maximizing its posterior probability yields a similar construction to a Variational Auto-Encoder (VAE), but with a structured prior coupled by non-linear equations. This prior represents an interpretable SEM with learnable parameters (such as a physical model or dependence structure), which can be fitted to data while simultaneously learning the latent variables. Unfortunately, computing KL-divergences with this non-linear prior is intractable. We show how linearizing arbitrary SEMs via back-propagation produces local non-isotropic Gaussian priors, for which the KL-divergences can be computed efficiently and differentiably. We propose two versions, one for IID data (such as images) which detects related causal variables within a sample, and one for non-IID data (such as video) which detects variables that are also related over time. Our proposal is complementary to causal discovery techniques, which assume given variables, and instead discovers both variables and their causal relationships. We experiment with recovering causal models from images, and learning temporal relations based on the Super Mario Bros videogame.

****

The Biased Artist: Exploiting Cultural Biases via Homoglyphs in Text-Guided Image Generation Models

Lukas Struppek,Dominik Hintersdorf,Kristian Kersting

Text-guided image generation models, such as DALL-E2 and Stable Diffusion, have recently received much attention from academia and the general public. Provided with textual descriptions, these models are capable of generating high-quality images depicting various concepts and styles. However, such models are trained on large amounts of public data and implicitly learn relationships from their training data that are not immediately apparent. We demonstrate that common multimodal models implicitly learned cultural biases that can be triggered and injected into the generated images by simply replacing single characters in the textual d escription with visually similar non-Latin characters. These so-called homoglyph replacements enable malicious users or service providers to induce biases into the generated images and even render the whole generation process useless. We practically illustrate such attacks on DALL-E2 and Stable Diffusion as text-guided image generation models and further show that CLIP also behaves similarly. Our results further indicate that text encoders trained on multilingual data provide a way to mitigate the effects of homoglyph replacements.

****

Parameterized projected Bellman operator

Théo Vincent,Alberto Maria Metelli,Jan Peters,Marcello Restelli,Carlo D'Eramo

The Bellman operator is a cornerstone of reinforcement learning, widely used in a plethora of works, from value-based methods to modern actor-critic approaches. In problems with unknown models, the Bellman operator requires transition samples that strongly determine its behavior, as uninformative samples can result in negligible updates or long detours before reaching the fixed point. In this work, we introduce the novel idea of obtaining an approximation of the Bellman operator, which we call projected Bellman operator (PBO). Our PBO is a parametric operator on the parameter space of a given value function. Given the parameters of a value function, PBO outputs the parameters of a new value function and converges to a fixed point in the limit, as a standard Bellman operator. Notably, our PBO can approximate repeated applications of the true Bellman operator at once, as opposed to the sequential nature of the standard Bellman operator. We prove the important consequences of this finding for different classes of problems by analyzing PBO in terms of stability, convergence, and approximation error. Eventually, we propose an approximate value-iteration algorithm to show how PBO can overcome the limitations of classical methods, opening up multiple research directions as a novel paradigm in reinforcement learning.

****

Cross-Level Distillation and Feature Denoising for Cross-Domain Few-Shot Classification

Hao ZHENG,Runqi Wang,Jianzhuang Liu,Asako Kanezaki

The conventional few-shot classification aims at learning a model on a large labeled base dataset and rapidly adapting to a target dataset that is from the same distribution as the base dataset. However, in practice, the base and the target datasets of few-shot classification are usually from different domains, which is the problem of cross-domain few-shot classification. We tackle this problem by making a small proportion of unlabeled images in the target domain accessible in the training stage. In this setup, even though the base data are sufficient and labeled, the large domain shift still makes transferring the knowledge from the base dataset difficult. We meticulously design a cross-level knowledge distillation method, which can strengthen the ability of the model to extract more discriminative features in the target dataset by guiding the network's shallow layers to learn higher-level information. Furthermore, in order to alleviate the overfitting in the evaluation stage, we propose a feature denoising operation which can reduce the feature redundancy and mitigate overfitting. Our approach can surpass the previous state-of-the-art method, Dynamic-Distillation, by 5.44% on 1-shot and 1.37% on 5-shot classification tasks on average in the BSCD-FSL benchmark. The implementation code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/CLDFD.

****

kaBEDONN: posthoc eXplainable Artificial Intelligence with Data Ordered Neural Network

Erico Tjoa,Cuntai Guan

Different approaches to eXplainable Artificial Intelligence (XAI) have been explored including (1) the systematic study of the effect of individual training data sample on the final model (2) posthoc attribution methods that assign importance values to the components of each data sample. Combining concepts from both approaches, we introduce kaBEDONN, a system of ordered dataset coupled with a posthoc and model-agnostic method for querying \textit{relevant} training data samples. These \textit{relevant} data are intended as the explanations for model pred

ictions that are both user-friendly and easily adjustable by developers. Explanations can thus be finetuned and damage control can be performed with ease.
****

DELTA: DEGRADATION-FREE FULLY TEST-TIME ADAPTATION
Bowen Zhao,Chen Chen,Shu-Tao Xia
Fully test-time adaptation aims at adapting a pre-trained model to the test stream during real-time inference, which is urgently required when the test distribution differs from the training distribution. Several efforts have been devoted to improving adaptation performance. However, we find that two unfavorable defects are concealed in the prevalent adaptation methodologies like test-time batch normalization (BN) and self-learning. First, we reveal that the normalization statistics in test-time BN are completely affected by the currently received test samples, resulting in inaccurate estimates. Second, we show that during test-time adaptation, the parameter update is biased towards some dominant classes. In addition to the extensively studied test stream with independent and class-balanced samples, we further observe that the defects can be exacerbated in more complicated test environments, such as (time) dependent or class-imbalanced data. We observe that previous approaches work well in certain scenarios while show performance degradation in others due to their faults. In this paper, we provide a plug-in solution called DELTA for Degradation-freE fuLly Test-time Adaptation, which consists of two components: (i) Test-time Batch Renormalization (TBR), introduced to improve the estimated normalization statistics. (ii) Dynamic Online re-weighTing (DOT), designed to address the class bias within optimization. We investigate various test-time adaptation methods on three commonly used datasets with four scenarios, and a newly introduced real-world dataset. DELTA can help them deal with all scenarios simultaneously, leading to SOTA performance.
****

Bit-Pruning: A Sparse Multiplication-Less Dot-Product
Yusuke Sekikawa,Shingo Yashima
Dot-product is a central building block in neural networks.
However, multiplication ($\texttt{mult}$) in dot-product consumes intensive energy and space costs that challenge deployment on resource-constrained edge devices.
In this study, we realize energy-efficient neural networks by exploiting a $\texttt{mult}$-less, sparse dot-product. We first reformulate a dot-product between an integer weight and activation into an equivalent operation comprised of additions followed by bit-shifts ($\texttt{add-shift-add}$).
In this formulation, the number of $\texttt{add}$ operations equals the number of bits of the integer weight in binary format.
Leveraging this observation, we propose Bit-Pruning, which removes unnecessary bits in each weight value during training to reduce the energy consumption of $\texttt{add-shift-add}$. Bit-Pruning can be seen as soft Weight-Pruning as it prunes bits, not the whole weight element.
In extensive experiments, we demonstrate that sparse $\texttt{mult}$-less networks trained with Bit-Pruning show a better accuracy-energy trade-off than sparse $\texttt{mult}$ networks trained with Weight-Pruning.
****

Abstract-to-Executable Trajectory Translation for One-Shot Task Generalization
Stone Tao,Xiaochen Li,Tongzhou Mu,Zhiao Huang,Yuzhe Qin,Hao Su
Training long-horizon robotic policies in complex physical environments is essential for many applications, such as robotic manipulation. However, learning a policy that can generalize to unseen tasks is challenging. In this work, we propose to achieve one-shot task generalization by decoupling plan generation and plan execution. Specifically, our method solves complex long-horizon tasks in three steps: build a paired abstract environment by simplifying geometry and physics, generate abstract trajectories, and solve the original task by an abstract-to-executable trajectory translator. In the abstract environment, complex dynamics such as physical manipulation are removed, making abstract trajectories easier to generate. However, this introduces a large domain gap between abstract trajectories and the actual executed trajectories as abstract trajectories lack low-level

details and aren’t aligned frame-to-frame with the executed trajectory. In a manner reminiscent of language translation, our approach leverages a seq-to-seq model to overcome the large domain gap between the abstract and executable trajectories, enabling the low-level policy to follow the abstract trajectory. Experimental results on various unseen long-horizon tasks with different robot embodiments demonstrate the practicability of our methods to achieve one-shot task generalization. Videos and more details can be found in the supplementary materials and project page: https://sites.google.com/view/abstract-to-executable-iclr23/

****

Unveiling The Mask of Position-Information Pattern Through the Mist of Image Features

Chieh Hubert Lin,Hung-Yu Tseng,Hsin-Ying Lee,Maneesh Kumar Singh,Ming-Hsuan Yang

Recent studies have shown that paddings in convolutional neural networks encode absolute position information which can negatively affect the model performance for certain tasks. However, existing metrics for quantifying the strength of positional information remain unreliable and frequently lead to erroneous results. To address this issue, we propose novel metrics for measuring and visualizing the encoded positional information. We formally define the encoded information as Position-information Pattern from Padding (PPP) and conduct a series of experiments to study its properties as well as its formation. The proposed metrics measure the presence of positional information more reliably than the existing metrics based on PosENet and tests in F-Conv. We also demonstrate that for any extant (and proposed) padding schemes, PPP is primarily a learning artifact and is less dependent on the characteristics of the underlying padding schemes.

****

kNN-Diffusion: Image Generation via Large-Scale Retrieval

Shelly Sheynin,Oron Ashual,Adam Polyak,Uriel Singer,Oran Gafni,Eliya Nachmani,Yaniv Taigman

Recent text-to-image models have achieved impressive results. However, since they require large-scale datasets of text-image pairs, it is impractical to train them on new domains where data is scarce or not labeled.

In this work, we propose using large-scale retrieval methods, in particular, efficient k-Nearest-Neighbors (kNN), which offers novel capabilities: (1) training a substantially small and efficient text-to-image diffusion model using only pre-trained multi-modal embeddings, but without an explicit text-image dataset, (2) generating out-of-distribution images by simply swapping the retrieval database at inference time, and (3) performing text-driven local semantic manipulations while preserving object identity. To demonstrate the robustness of our method, we apply our kNN approach on two state-of-the-art diffusion backbones, and show results on several different datasets. As evaluated by human studies and automatic metrics, our method achieves state-of-the-art results compared to existing approaches that train text-to-image generation models using images-only dataset.

****

IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION?

Ruifei He,Shuyang Sun,Xin Yu,Chuhui Xue,Wenqing Zhang,Philip Torr,Song Bai,XIAOJUAN QI

Recent text-to-image generation models have shown promising results in generating high-fidelity photo-realistic images. Though the results are astonishing to human eyes, how applicable these generated images are for recognition tasks remains under-explored. In this work, we extensively study whether and how synthetic images generated from state-of-the-art text-to-image generation models can be used for image recognition tasks, and focus on two perspectives: synthetic data for improving classification models in the data-scare settings (i.e. zero-shot and few-shot), and synthetic data for large-scale model pre-training for transfer learning. We showcase the powerfulness and shortcomings of synthetic data from existing generative models, and propose strategies for better applying synthetic data for recognition tasks. Code: https://github.com/CVMI-Lab/SyntheticData.

****

Learnable Behavior Control: Breaking Atari Human World Records via Sample-Efficient Behavior Selection

Jiajun Fan,Yuzheng Zhuang,Yuecheng Liu,Jianye HAO,Bin Wang,Jiangcheng Zhu,Hao Wang,Shu-Tao Xia

The exploration problem is one of the main challenges in deep reinforcement learning (RL). Recent promising works tried to handle the problem with population-based methods, which collect samples with diverse behaviors derived from a population of different exploratory policies. Adaptive policy selection has been adopted for behavior control. However, the behavior selection space is largely limited by the predefined policy population, which further limits behavior diversity.

In this paper, we propose a general framework called Learnable Behavioral Control (LBC) to address the limitation, which a) enables a significantly enlarged behavior selection space via formulating a hybrid behavior mapping from all policies; b) constructs a unified learnable process for behavior selection. We introduce LBC into distributed off-policy actor-critic methods and achieve behavior control via optimizing the selection of the behavior mappings with bandit-based meta-controllers. Our agents have achieved 10077.52% mean human normalized score and surpassed 24 human world records within 1B training frames in the Arcade Learning Environment, which demonstrates our significant state-of-the-art (SOTA) performance without degrading the sample efficiency.

****

Decompose to Generalize: Species-Generalized Animal Pose Estimation

Guangrui Li,Yifan Sun,Zongxin Yang,Yi Yang

This paper challenges the cross-species generalization problem for animal pose estimation, aiming to learn a pose estimator that can be well generalized to novel species. We find the relation between different joints is important with two-fold impact: 1) on the one hand, some relation is consistent across all the species and may help two joints mutually confirm each other, e.g., the eyes help confirm the nose and vice versa because they are close in all species. 2) on the other hand, some relation is inconsistent for different species due to the species variation and may bring severe distraction rather than benefit. With these two insights, we propose a Decompose-to-Generalize (D-Gen) pose estimation method to break the inconsistent relations while preserving the consistent ones. Specifically, D-Gen first decomposes the body joints into several joint concepts so that each concept contains multiple closely-related joints. Given these joint concepts, D-Gen 1) promotes the interaction between intra-concept joints to enhance their reliable mutual confirmation, and 2) suppresses the interaction between inter-concept joints to prohibit their mutual distraction. Importantly, we explore various decomposition approaches, i.e., heuristic, geometric and attention-based approaches. Experimental results show that all these decomposition manners yield reasonable joint concepts and substantially improve cross-species generalization (and the attention-based approach is the best).

****

IDEAL: Query-Efficient Data-Free Learning from Black-Box Models

Jie Zhang,Chen Chen,Lingjuan Lyu

Knowledge Distillation (KD) is a typical method for training a lightweight student model with the help of a well-trained teacher model.

However, most KD methods require access to either the teacher's training data or model parameter, which is unrealistic. To tackle this problem, recent works study KD under data-free and black-box settings. Nevertheless, these works require a large number of queries to the teacher model, which incurs significant monetary and computational costs. To address these problems, we propose a novel method called \emph{query-effIcient Data-free lEarning from blAck-box modeLs} (IDEAL), which aims to query-efficiently learn from black-box model APIs to train a good student without any real data.   In detail, IDEAL trains the student model in two stages: data generation and model distillation. Note that IDEAL does not require any query in the data generation stage and queries the teacher only once for each sample in the distillation stage. Extensive experiments on various real-world datasets show the effectiveness of the proposed IDEAL. For instance, IDEAL can improve the performance of the best baseline method DFME by 5.83\% on CIFAR10 dataset with only $0.02\times$ the query budget of DFME.

****

MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction

Bencheng Liao,Shaoyu Chen,Xinggang Wang,Tianheng Cheng,Qian Zhang,Wenyu Liu,Chang Huang

High-definition (HD) map provides abundant and precise environmental information of the driving scene, serving as a fundamental and indispensable component for planning in autonomous driving system. We present MapTR, a structured end-to-end Transformer for efficient online vectorized HD map construction. We propose a unified permutation-equivalent modeling approach, i.e., modeling map element as a point set with a group of equivalent permutations, which accurately describes the shape of map element and stabilizes the learning process. We design a hierarchical query embedding scheme to flexibly encode structured map information and perform hierarchical bipartite matching for map element learning. MapTR achieves the best performance and efficiency with only camera input among existing vectorized map construction approaches on nuScenes dataset. In particular, MapTR-nano runs at real-time inference speed (25.1 FPS) on RTX 3090, $8\times$ faster than the existing state-of-the-art camera-based method while achieving 5.0 higher mAP. Even compared with the existing state-of-the-art multi-modality method, MapTR-nano achieves 0.7 higher mAP and $8\times$ faster inference speed, and MapTR-tiny achieves 13.5 higher mAP and $3\times$ faster inference speed. Abundant qualitative results show that MapTR maintains stable and robust map construction quality in complex and various driving scenes. MapTR is of great application value in autonomous driving. Code and more demos are available at https://github.com/hustvl/MapTR.

****

(LA)YER-NEIGH(BOR) SAMPLING: DEFUSING NEIGHBORHOOD EXPLOSION

Muhammed Fatih Balin,Umit Catalyurek

Graph Neural Networks have recently received a significant attention, however, training them at a large scale still remains as a challenge.
Minibatch training coupled with sampling is used to alleviate this challenge.
However existing approaches either suffer from the neighborhood explosion phenomenon or does not have good performance.
To deal with these issues, we propose a new sampling algorithm called LAyer-neighBOR sampling (LABOR).
It is designed to be a direct replacement for Neighborhood Sampling with the same fanout hyperparameter while sampling much fewer vertices, without sacrificing quality.
By design, the variance of the estimator of each vertex matches Neighbor Sampling from the point of view from a single vertex.
In our experiments, we demonstrate the superiority of our approach when it comes to model convergence behaviour against Neighbor Sampling and also the other Layer Sampling approaches under the same limited vertex sampling budget constraints.

****

Probing into Overfitting for Video Recognition

Yitian Zhang,Yue Bai,Huan Wang,Yizhou Wang,Yun Fu

Video recognition methods based on 2D networks have thrived in recent years, leveraging advanced image classification techniques. However, overfitting is an even severe problem in 2D video recognition models as 1) the scale of video datasets is relatively small compared to image recognition datasets like ImageNet; 2) current pipeline treats background and semantic frames equally during optimization which aggravates overfitting. Based on these challenges, we design a video-specific data augmentation approach, named as Ghost Motion (GM), to alleviate overfitting. Specifically, GM shifts channels along temporal dimension to enable semantic motion information diffused into other frames which may be irrelevant originally, leading to improvement in frame-wise accuracy. In addition, for challenging video samples with significant temporal dependency (e.g., Something-Something), we further scale the logits during training to prevent overconfident predictions on background frames. Comprehensive empirical validation on various popular datasets shows that the proposed method can improve the generalization of existi ng methods and is compatible to other competing data augmentation approaches.

****

Image as Set of Points

Xu Ma,Yuqian Zhou,Huan Wang,Can Qin,Bin Sun,Chang Liu,Yun Fu

What is an image, and how to extract latent features?

Convolutional Networks (ConvNets) consider an image as organized pixels in a rectangular shape and extract features via convolutional operation in a local region; Vision Transformers (ViTs) treat an image as a sequence of patches and extract features via attention mechanism in a global range. In this work, we introduce a straightforward and promising paradigm for visual representation, which is called Context Clusters. Context clusters (CoCs) view an image as a set of unorganized points and extract features via a simplified clustering algorithm. In detail, each point includes the raw feature (e.g., color) and positional information (e.g., coordinates), and a simplified clustering algorithm is employed to group and extract deep features hierarchically. Our CoCs are convolution- and attention-free, only relying on clustering algorithm for spatial interaction. Owing to the simple design, we show CoCs endow gratifying interpretability via the visualization of the clustering process.

Our CoCs aim at providing a new perspective on image and visual representation, which may enjoy broad applications in different domains and exhibit profound insights. Even though we are not targeting SOTA performance, COCs still achieve comparable or even better performance than ConvNets or ViTs on several benchmarks.

****

Hybrid Neuro-Symbolic Reasoning based on Multimodal Fusion

Subrata Das,Bodong Zhou

Deep neural models and symbolic Artificial Intelligence (AI) systems have contrasting advantages and disadvantages. Neural models can be trained from raw, incomplete and noisy data to obtain abstraction of features at various levels, but their uninterpretability is well-known. On the other hand, the traditional rule-based symbolic reasoning encodes domain knowledge, but its failure is often attributed to the acquisition bottleneck. We propose to build a hybrid learning and reasoning system which is based on multimodal fusion approach that brings together advantageous features from both the paradigms. Specifically, we enhance convolutional neural networks (CNNs) with the structured information of ‘if-then’ symbolic logic rules obtained via word embeddings corresponding to propositional symbols and terms. With many dozens of intuitive rules relating the type of a scene with its typical constituent objects, we are able to achieve significant improvement over the base CNN-based classification. Our approach is extendible to handle first-order logical syntax for rules and other deep learning models.

****

Distilling Text-Image Foundation Models

Qifeng Wu,Huan Wang,Xu Ma,Yun Fu

Large pretrained foundation models (such as CLIP, DALL-E) are among the most recent significant advances in the AI community. Their implication is profound. This paper examines the value of these foundation models as a model knowledge base -- we aim to distill the knowledge in these foundation models for training lightweight models designed for specific tasks in practical application scenarios with improved performance. Despite abundant progress in knowledge distillation (KD) in traditional models trained under the supervision of class labels in datasets encoded as integers, distilling such text-image contrastive learning model has not been explored extensively. Meanwhile, KD is well-known for being bothered by the capacity gap problem (i.e., distilling knowledge from a teacher significantly larger than a student often degrades the performance of the student). The teacher-student capacity gap in distilling foundation models is even larger. Therefore, how to overcome this potential issue is also elusive now. This paper presents detailed analyses of these questions aiming to successfully tap into a pretrained foundation model (CLIP) to boost the student's performance. Besides the practical performance benefits, several interesting discoveries are unveiled: (1) CLIP is not bothered by the capacity gap, which may let us re-evaluate if the "ca

pacity-gap" issue is really due to the capacity gap (2) We find the reason is largely due to that CLIP is not over-confident on the wrong labels when misclassifies input image samples.
****
Trainability Preserving Neural Pruning
Huan Wang,Yun Fu
Many recent works have shown trainability plays a central role in neural network pruning -- unattended broken trainability can lead to severe under-performance and unintentionally amplify the effect of retraining learning rate, resulting in biased (or even misinterpreted) benchmark results. This paper introduces trainability preserving pruning (TPP), a scalable method to preserve network trainability against pruning, aiming for improved pruning performance and being more robust to retraining hyper-parameters (e.g., learning rate). Specifically, we propose to penalize the gram matrix of convolutional filters to decorrelate the pruned filters from the retained filters. In addition to the convolutional layers, per the spirit of preserving the trainability of the whole network, we also propose to regularize the batch normalization parameters (scale and bias). Empirical studies on linear MLP networks show that TPP can perform on par with the oracle trainability recovery scheme. On nonlinear ConvNets (ResNet56/VGG19) on CIFAR10/100, TPP outperforms the other counterpart approaches by an obvious margin. Moreover, results on ImageNet-1K with ResNets suggest that TPP consistently performs more favorably against other top-performing structured pruning approaches. Code: https://github.com/MingSun-Tse/TPP.
****
Robustness Exploration of Semantic Information in Adversarial Training
Huafeng Kuang,Hong Liu,Mingliang Xu,YONGJIAN WU,Rongrong Ji
In this paper, we look into the problem of adversarial robustness from the semantic information perspective. We demonstrate a novel insight that adversarial attacks destroy the correlation between visual representations and semantic word vectors, and adversarial training fixed it. We further find that the correlation between robust features of different categories is consistent with the correlation between corresponding semantic word vectors. Based on that, we introduce the semantic information to assist model training and propose Semantic Constraint Adversarial Robust Learning (SCARL). First, we follow an information-theoretical lens to formulate the mutual information between the visual representation and the corresponding semantic word vector in the embedding space to bridge the information gap. We further provide a differentiable lower bound to optimize such mutual information efficiently. Second, we propose a novel semantic structural constraint, encouraging the trained model to keep the structure of visual representations consistent with that of semantic word vectors. Finally, we combine these two techniques with adversarial training to learn robust visual representation. Experimentally, we conduct extensive experiments on several benchmarks, demonstrating that semantic information is indeed beneficial to model robustness.
****
Learning Implicit Scale Conditioned Memory Compensation for Talking Head Generation
Fa-Ting Hong,Dan Xu
Talking head video generation aims to animate the pose and expression of a person in a target driving video using motion information contained in the video, while maintaining a person's identity in a given still source image. Highly dynamic and complex motions in the driving video cause ambiguous generation from the source image, because the still source image cannot provide sufficient appearance information for occluded regions or delicate expressions, which severely produces artifacts and significantly degrades the generation quality. However, existing works mainly focus on learning more accurate motion estimation and representation in 2D and 3D, and they ignore the facial structural prior in addressing the facial ambiguities. Therefore, effective handling of the ambiguities in the dramatic appearance changes of the source to largely improve facial details and completeness in generation still remains barely explored. To this end, we propose a novel implicit scale conditioned memory compensation network (MCNet) for high-fid

elity talking head generation. Specifically, considering human faces are symmetric and structured, we aim to automatically learn a representative global facial memory bank from all training data as a prior to compensate for the facial generation features. Each face in the source image contains a scale that can be reflected in detected facial keypoints. To better query the learned global memory, we further propose to learn implicit scale representations from the discrete keypoints, which can be used to condition on the query of the global memory, to obtain scale-aware memory for the feature compensation. Extensive experiments from quantitative and qualitative perspectives demonstrate that MCNet can learn representative and complementary facial memory, and can clearly outperform previous state-of-the-art methods on VoxCeleb1 and CelebV datasets.

****

Diagnosing and Rectifying Vision Models using Language

Yuhui Zhang,Jeff Z. HaoChen,Shih-Cheng Huang,Kuan-Chieh Wang,James Zou,Serena Yeung

Recent multi-modal contrastive learning models have demonstrated the ability to learn an embedding space suitable for building strong vision classifiers, by leveraging the rich information in large-scale image-caption datasets. Our work highlights a distinct advantage of this multi-modal embedding space: the ability to diagnose vision classifiers through natural language. The traditional process of diagnosing model behaviors in deployment settings involves labor-intensive data acquisition and annotation. Our proposed method can discover high-error data slices, identify influential attributes and further rectify undesirable model behaviors, without requiring any visual data. Through a combination of theoretical explanation and empirical verification, we present conditions under which classifiers trained on embeddings from one modality can be equivalently applied to embeddings from another modality. On a range of image datasets with known error slices, we demonstrate that our method can effectively identify the error slices and influential attributes, and can further use language to rectify failure modes of the classifier.

****

Harnessing Out-Of-Distribution Examples via Augmenting Content and Style

Zhuo Huang,Xiaobo Xia,Li Shen,Bo Han,Mingming Gong,Chen Gong,Tongliang Liu

Machine learning models are vulnerable to Out-Of-Distribution (OOD) examples, such a problem has drawn much attention. However, current methods lack a full understanding of different types of OOD data: there are benign OOD data that can be properly adapted to enhance the learning performance, while other malign OOD data would severely degenerate the classification result. To Harness OOD data, this paper proposes HOOD method that can leverage the content and style from each image instance to identify benign and malign OOD data. Particularly, we design a variational inference framework to causally disentangle content and style features by constructing a structural causal model. Subsequently, we augment the content and style through an intervention process to produce malign and benign OOD data, respectively. The benign OOD data contain novel styles but hold our interested contents, and they can be leveraged to help train a style-invariant model. In contrast, the malign OOD data inherit unknown contents but carry familiar styles, by detecting them can improve model robustness against deceiving anomalies. Thanks to the proposed novel disentanglement and data augmentation techniques, HOOD can effectively deal with OOD examples in unknown and open environments, whose effectiveness is empirically validated in three typical OOD applications including OOD detection, open-set semi-supervised learning, and open-set domain adaptation.

****

On Stability and Generalization of Bilevel Optimization Problems

Meng Ding,Mingxi Lei,Yunwen Lei,Di Wang,Jinhui Xu

(Stochastic) bilevel optimization is a frequently encountered problem in machine learning with a wide range of applications such as meta-learning, hyper-parameter optimization, and reinforcement learning. Most of the existing studies on this problem only focused on analyzing the convergence or improving the convergence rate, while little effort has been devoted to understanding its generalization behaviors. In this paper, we conduct a thorough analysis on the generalization of first-order (gradient-based) methods for the bilevel optimization problem. We first establish a fundamental connection between algorithmic stability and generalization error in different forms and give a high probability generalization bound which improves the previous best one from $O(\sqrt{n})$ to $O(\log n)$, where n is the sample size. We then provide the first stability bounds for the general case where both inner and outer level parameters are subject to continuous update, while existing work allows only the outer level parameter to be updated. Our analysis can be applied in various standard settings such as strongly-convex-strongly-convex (SC-SC), convex-convex (C-C), and nonconvex-nonconvex (NC-NC). Our analysis for the NC-NC setting can also be extended to a particular nonconvex-strongly-convex (NC-SC) setting that is commonly encountered in practice. Finally, we corroborate our theoretical analysis and demonstrate how iterations can affect the generalization error by experiments on meta-learning and hyper-parameter optimization.

****

Learning GFlowNets from partial episodes for improved convergence and stability

Kanika Madan,Jarrid Rector-Brooks,Maksym Korablyov,Emmanuel Bengio,Moksh Jain,Andrei Cristian Nica,Tom Bosc,Yoshua Bengio,Nikolay Malkin

Generative flow networks (GFlowNets) are a family of algorithms for training a sequential sampler of discrete objects under an unnormalized target density and have been successfully used for various probabilistic modeling tasks. Existing training objectives for GFlowNets are either local to states or transitions, or propagate a reward signal over an entire sampling trajectory. We argue that these alternatives represent opposite ends of a gradient bias-variance tradeoff and propose a way to exploit this tradeoff to mitigate its harmful effects. Inspired by the TD(λ) algorithm in reinforcement learning, we introduce subtrajectory balance or SubTB(λ), a GFlowNet training objective that can learn from partial action subsequences of varying lengths. We show that SubTB(λ) accelerates sampler convergence in previously studied and new environments and enables training GFlowNets in environments with longer action sequences and sparser reward landscapes than what was possible before. We also perform a comparative analysis of stochastic gradient dynamics, shedding light on the bias-variance tradeoff in GFlowNet training and the advantages of subtrajectory balance.

****

DropIT: Dropping Intermediate Tensors for Memory-Efficient DNN Training

Joya Chen,Kai Xu,Yuhui Wang,Yifei Cheng,Angela Yao

A standard hardware bottleneck when training deep neural networks is GPU memory. The bulk of memory is occupied by caching intermediate tensors for gradient computation in the backward pass. We propose a novel method to reduce this footprint - Dropping Intermediate Tensors (DropIT). DropIT drops min-k elements of the intermediate tensors and approximates gradients from the sparsified tensors in the backward pass. Theoretically, DropIT reduces noise on estimated gradients and therefore has a higher rate of convergence than vanilla-SGD. Experiments show that we can drop up to 90\% of the intermediate tensor elements in fully-connected and convolutional layers while achieving higher testing accuracy for Visual Transformers and Convolutional Neural Networks on various tasks (e.g., classification, object detection, instance segmentation). Our code and models are available at https://github.com/chenjoya/dropit.

****

Self-attentive Rationalization for Graph Contrastive Learning

Sihang Li,Yanchen Luo,An Zhang,Xiang Wang,Xiangnan He,Tat-Seng Chua

Graph augmentation is the key component to reveal instance-discriminative features of a graph as its rationale in graph contrastive learning (GCL).

And existing rationale-aware augmentation mechanisms in GCL frameworks roughly fall into two categories and suffer from inherent limitations: (1) non-heuristic methods with the guidance of domain knowledge to preserve salient features, which require expensive expertise and lacks generality, or (2) heuristic augmentations with a co-trained auxiliary model to identify crucial substructures, which face not only the dilemma between system complexity and transformation diversity, but also the instability stemming from the co-training of two separated sub-models.

Inspired by recent studies on transformers, we propose $\underline{S}$elf-attentive $\underline{R}$ationale guided $\underline{G}$raph $\underline{C}$ontrastive $\underline{L}$earning (SR-GCL), which integrates rationale finder and encoder together, leverages the self-attention values in transformer module as a natural guidance to delineate semantically informative substructures from both node- and edge-wise views, and contrasts on rationale-aware augmented pairs.

On real world biochemistry datasets, visualization results verify the effectiveness of self-attentive rationalization and the performance on downstream tasks demonstrates the state-of-the-art performance of SR-GCL for graph model pre-training.

****

A Unified Framework for Soft Threshold Pruning

Yanqi Chen,Zhengyu Ma,Wei Fang,Xiawu Zheng,Zhaofei Yu,Yonghong Tian

Soft threshold pruning is among the cutting-edge pruning methods with state-of-the-art performance. However, previous methods either perform aimless searching on the threshold scheduler or simply set the threshold trainable, lacking theoretical explanation from a unified perspective. In this work, we reformulate soft threshold pruning as an implicit optimization problem solved using the Iterative Shrinkage-Thresholding Algorithm (ISTA), a classic method from the fields of sparse recovery and compressed sensing. Under this theoretical framework, all threshold tuning strategies proposed in previous studies of soft threshold pruning are concluded as different styles of tuning L_1-regularization term. We further derive an optimal threshold scheduler through an in-depth study of threshold scheduling based on our framework. This scheduler keeps L_1-regularization coefficient stable, implying a time-invariant objective function from the perspective of optimization. In principle, the derived pruning algorithm could sparsify any mathematical model trained via SGD. We conduct extensive experiments and verify its state-of-the-art performance on both Artificial Neural Networks (ResNet-50 and MobileNet-V1) and Spiking Neural Networks (SEW ResNet-18) on ImageNet datasets. On the basis of this framework, we derive a family of pruning methods, including sparsify-during-training, early pruning, and pruning at initialization. The code is available at https://github.com/Yanqi-Chen/LATS.

****

Efficient Automatic Machine Learning via Design Graphs

Ying-Xin Wu,Jiaxuan You,Jure Leskovec,Zhitao Ying

Despite the success of automated machine learning (AutoML), which aims to find the best design, including the architecture of deep networks and hyper-parameters, conventional AutoML methods are computationally expensive and hardly provide insights into the relations of different model design choices. To tackle the challenges, we propose FALCON, an efficient sample-based method to search for the optimal model design. Our key insight is to model the design space of possible model designs as a design graph, where the nodes represent design choices, and the edges denote design similarities. FALCON features 1) a task-agnostic module, which performs message passing on the design graph via a Graph Neural Network (GNN), and 2) a task-specific module, which conducts label propagation of the known model performance information on the design graph. Both modules are combined to predict the design performances in the design space, navigating the search direction. We conduct extensive experiments on 27 node and graph classification tasks from various application domains, and an image classification task on the CIFAR-10 dataset. We empirically show that FALCON can efficiently obtain the well-performing designs for each task using only 30 explored nodes. Specifically, FALCON has a comparable time cost with the one-shot approaches while achieving an average improvement of 3.3% compared with the best baselines.

****

TaskPrompter: Spatial-Channel Multi-Task Prompting for Dense Scene Understanding

Hanrong Ye,Dan Xu

Learning effective representations simultaneously from multiple tasks in a unified network framework is a fundamental paradigm for multi-task dense visual scene understanding. This requires joint modeling (i) task-generic and (ii) task-specific representations, and (iii) cross-task representation interactions. Existing works typically model these three perspectives with separately designed structures, using shared network modules for task-generic learning, different modules for task-specific learning, and establishing connections among these components for cross-task interactions. It is barely explored in the literature to model these three perspectives in each network layer in an end-to-end manner, which can not only minimize the effort of carefully designing empirical structures for the three multi-task representation learning objectives, but also greatly improve the representation learning capability of the multi-task network since all the model capacity will be used to optimize the three objectives together. In this paper, we propose TaskPrompter, a novel spatial-channel multi-task prompting transformer framework to achieve this target. Specifically, we design a set of spatial-channel task prompts and learn their spatial- and channel interactions with the shared image tokens in each transformer layer with attention mechanism, as aggregating spatial and channel information is critical for dense prediction tasks. Each task prompt learns task-specific representation for one task, while all the prompts can jointly contribute to the learning of the shared image token representations, and the interactions between different task prompts model the cross-task relationship. To decode dense predictions for multiple tasks with the learned spatial-channel task prompts from transformer, we accordingly design a dense task prompt decoding mechanism, which queries the shared image tokens using task prompts to obtain spatial- and channel-wise task-specific representations. Extensive experiments on two challenging multi-task dense scene understanding benchmarks (i.e. NYUD-V2 and PASCAL-Context) show the superiority of the proposed framework and TaskPrompter establishes significant state-of-the-art performances on multi-task dense predictions. Codes and models are made publicly available at https://github.com/prismformore/Multi-Task-Transformer.

****

Individual Privacy Accounting for Differentially Private Stochastic Gradient Descent

Da Yu,Gautam Kamath,Janardhan Kulkarni,Tie-Yan Liu,Jian Yin,Huishuai Zhang

Differentially private stochastic gradient descent (DP-SGD) is the workhorse algorithm for recent advances in private deep learning. It provides a single privacy guarantee to all datapoints in the dataset. We propose an efficient algorithm to compute privacy guarantees for individual examples when releasing models trained by DP-SGD. We use our algorithm to investigate individual privacy parameters across a number of datasets. We find that most examples enjoy stronger privacy guarantees than the worst-case bound. We further discover that the training loss and the privacy parameter of an example are well-correlated. This implies groups that are underserved in terms of model utility are simultaneously underserved in terms of privacy guarantee. For example, on CIFAR-10, the average ε of the class with the lowest test accuracy is 43.6% higher than that of the class with the highest accuracy. We also run membership inference attacks to show this reflects disparate empirical privacy risks.

****

CI-VAE: a Class-Informed Deep Variational Autoencoder for Enhanced Class-Specific Data Interpolation

Mohsen Nabian,Zahra Eftekhari,Alec Wong

We proposed Class-Informed Variational Autoencoder (CI-VAE) to enable interpolation between arbitrary pairs of observations of the same class. CI-VAE combines the general VAE architecture with a linear discriminator layer on the latent space to enforce the construction of a latent space such that observations from different classes are linearly separable. In conventional VAEs, class overlapping on the latent space usually occurs. However, in CI-VAE, the enforced linear separability of classes on the latent space allows for robust latent-space linear traversal and data generation between two arbitrary observations of the same class. Class-specific data interpolation has extensive potential applications in science, particularly in biology, such as uncovering the biological trajectory of dise ases or cancer. We used the MNIST dataset of handwritten digits as a case study to compare the performance of CI-VAE and VAE in class-specific data augmentation. We showed that CI-VAE significantly improved class-specific linear traversal and data augmentation compared with VAE while maintaining comparable reconstruction error. In a study of Colon cancer genomics data, we showed that the interpolation between normal cells and tumor cells using CI-VAE may enhance our understanding of cancer development.

****

Attention De-sparsification Matters: Inducing Diversity in Digital Pathology Representation Learning

Saarthak Kapse,Srijan Das,Jingwei Zhang,Rajarsi R. Gupta,Joel Saltz,Dimitris Samaras,Prateek Prasanna

In this work, we develop Di-SSL, a Diversity-inducing Self-Supervised Learning technique for histopathology image analysis. SSL techniques, such as contrastive and non-contrastive approaches, have been shown to learn rich and effective representations without any human supervision. Lately, computational pathology has also benefited from the resounding success of SSL. In this work, we develop a novel domain-aware pretext task to enhance representation learning in digital pathology. Our analysis of vanilla SSL-pretrained models’ attention distribution reveals an insightful observation: sparsity in attention, i.e, models tends to localize most of their attention to some prominent patterns in the image. Although atten- tion sparsity can be beneficial in natural images due to these prominent patterns being the object of interest itself, this can be sub-optimal in digital pathology; this is because, unlike natural images, digital pathology scans are not object-centric, but rather a complex phenotype of various spatially intermixed biological com- ponents. Inadequate diversification of attention in these complex images could result in crucial information loss. To address this, we first leverage cell segmenta- tion to densely extract multiple histopathology-specific representations. We then propose a dense pretext task for SSL, designed to match the multiple correspond- ing representations between the views. Through this, the model learns to attend to various components more closely and evenly, thus inducing adequate diversi- fication in attention for capturing context rich representations. Through quantita- tive and qualitative analysis on multiple slide-level tasks across cancer types, and patch-level classification tasks, we demonstrate the efficacy of our method and observe that the attention is more globally distributed. Specifically, we obtain a relative improvement in accuracy of up to 6.9% in slide-level and 2% in patch level classification tasks (corresponding AUC improvement up to 7.9% and 0.7%, respectively) over a baseline SSL model.

****

Learning Domain-Agnostic Representation for Disease Diagnosis

Churan Wang,Jing Li,Xinwei Sun,Fandong Zhang,Yizhou Yu,Yizhou Wang

In clinical environments, image-based diagnosis is desired to achieve robustness on multi-center samples. Toward this goal, a natural way is to capture only clinically disease-related features. However, such disease-related features are often entangled with center-effect, disabling robust transferring to unseen centers/domains. To disentangle disease-related features, we first leverage structural causal modeling to explicitly model disease-related and center-effects that are provable to be disentangled from each other. Guided by this, we propose a novel Domain Agnostic Representation Model (DarMo) based on variational Auto-Encoder. To facilitate disentanglement, we design domain-agnostic and domain-aware encoders to respectively capture disease-related features and varied center-effects by incorporating a domain-aware batch normalization layer. Besides, we constrain the disease-related features to well predict the disease label as well as clinical attributes, by leveraging Graph Convolutional Network (GCN) into our decoder. The effectiveness and utility of our method are demonstrated by the superior performance over others on both public datasets and inhouse datasets.

****

Boosting Out-of-Distribution Detection with Multiple Pre-trained Models

Feng Xue,Zi He,Chuanlong Xie,Falong Tan,Zhenguo Li

Out-of-Distribution (OOD) detection, i.e., identifying whether an input is sampl ed from a novel distribution other than the training distribution, is a critical task for safely deploying machine learning systems in the open world. Recently, post hoc detection utilizing pre-trained models has shown promising performance and can be scaled to large-scale problems. This advance raises a natural question: Can we leverage the diversity of multiple pre-trained models to improve the performance of post hoc detection methods? In this work, we propose a detection enhancement method by ensembling multiple detection decisions derived from a zoo of pre-trained models. Our approach uses the p-value instead of the commonly used hard threshold and leverages a fundamental framework of multiple hypothesis testing to control the true positive rate for In-Distribution (ID) data. We focus on the usage of model zoos and provide systematic empirical comparisons with current state-of-the-art methods on various OOD detection benchmarks. The proposed ensemble scheme shows consistent improvement compared to single-model detectors and significantly outperforms the current competitive methods. Our method substantially improves the relative performance by 65.40% and 26.96% on the CIFAR10 and ImageNet benchmarks.

****

Minimax Optimal Kernel Operator Learning via Multilevel Training

Jikai Jin,Yiping Lu,Jose Blanchet,Lexing Ying

Learning mappings between infinite-dimensional function spaces have achieved empirical success in many disciplines of machine learning, including generative modeling, functional data analysis, causal inference, and multi-agent reinforcement learning. In this paper, we study the statistical limit of learning a Hilbert-Schmidt operator between two infinite-dimensional Sobolev reproducing kernel Hilbert spaces. We establish the information-theoretic lower bound in terms of the Sobolev Hilbert-Schmidt norm and show that a regularization that learns the spectral components below the bias contour and ignores the ones above the variance contour can achieve the optimal learning rate. At the same time, the spectral components between the bias and variance contours give us flexibility in designing computationally feasible machine learning algorithms. Based on this observation, we develop a multilevel kernel operator learning algorithm that is optimal when learning linear operators between infinite-dimensional function spaces.

****

MixQuant: A Quantization Bit-width Search that Can Optimize the Performance of your Quantization Method

Eliska Kloberdanz,Wei Le

Quantization is a technique for creating efficient Deep Neural Networks (DNNs), which involves performing computations and storing tensors at lower bit-widths than f32 floating point precision. Quantization reduces model size and inference latency, and therefore allows for DNNs to be deployed on platforms with constrained computational resources and real-time systems. However, quantization can lead to numerical instability caused by roundoff error which leads to inaccurate computations and therefore, a decrease in quantized model accuracy. In this paper we focus on simulated quantized inference, where the quantized model parameters are stored in low-precision, but the mathematical operations on them (e.g. matrix multiplications and additions) are performed with floating point arithmetic. This means that the DNN parameters are first quantized from f32 to, for example, int4, and then dequantized back to f32 to perform computations. We show that the roundtrip process of quantizing and dequantizing the model parameters leads to roundoff error, which may lead to numerical instability. Similarly to prior works, which have shown that both biases and activations are more sensitive to quantization and are best kept in full precision or quantized with higher bit-widths, we show that some weights are more sensitive than others which should be reflected on their quantization bit-width. To that end we propose MixQuant, a search algorithm that finds the optimal custom quantization bit-width for each layer weight based on roundoff error and can be combined with any quantization method as a form of pre-processing optimization. We show that combining MixQuant with BRECQ, a state-of-the-art quantization method, yields better quantized model accuracy than BRECQ alone. Additionally, we combine MixQuant with vanilla asymmetric quantization to show that MixQuant has the potential to optimize the performance o f any quantization technique.

****

Logical Entity Representation in Knowledge-Graphs for Differentiable Rule Learning

Chi Han,Qizheng He,Charles Yu,Xinya Du,Hanghang Tong,Heng Ji

Probabilistic logical rule learning has shown great strength in logical rule mining and knowledge graph completion. It learns logical rules to predict missing edges by reasoning on existing edges in the knowledge graph. However, previous efforts have largely been limited to only modeling chain-like Horn clauses such as R1(x; z) ^ R2(z; y)) H(x; y). This formulation overlooks additional contextual information from neighboring sub-graphs of entity variables x, y and z. Intuitively, there is a large gap here, as local sub-graphs have been found to provide important information for knowledge graph completion. Inspired by these observations, we propose Logical Entity RePresentation (LERP) to encode contextual information of entities in the knowledge graph. A LERP is designed as a vector of probabilistic logical functions on the entity’s neighboring sub-graph. It is an interpretable representation while allowing for differentiable optimization. We can then incorporate LERP into probabilistic logical rule learning to learn more expressive rules. Empirical results demonstrate that with LERP, our model outperforms other rule learning methods in knowledge graph completion and is comparable or even superior to state-of-the-art black-box methods. Moreover, we find that our model can discover a more expressive family of logical rules. LERP can also be further combined with embedding learning methods like TransE to make it more interpretable.

****

S-SOLVER: Numerically Stable Adaptive Step Size Solver for Neural ODEs

Eliska Kloberdanz,Wei Le

A neural ordinary differential equation (ODE) is a relation between an unknown function and its derivatives, where the ODE is parameterized by a neural network. Therefore, to obtain a solution to a neural ODE requires a solver that performs numerical integration. Dopri5 is one of the most popular neural ODE solvers and also the default solver in torchdiffeq, a PyTorch library of ODE solvers. It is an adaptive step size solver based on the Runge-Kutta (RK) numerical methods. These methods rely on estimation of the local truncation error to select and adjust integration step size, which determines the numerical stability of the solution. A step size that is too large leads to numerical instability, while a step size that is too small may cause the solver to take unnecessarily many steps, which is computationally expensive and may even cause rounding error build up. Therefore, accurate local truncation error estimation is paramount for choosing an appropriate step size to obtain an accurate, numerically stable, and fast solution to the ODE. In this paper we propose a novel local truncation error approximation that is the first to consider solutions of four different RK orders to obtain a more reliable error estimate. This leads to a novel solver S-SOLVER (Stable Solver), which is more numerically stable; and therefore accurate. We demonstrate S-SOLVER's competitive performance in experiments on image recognition with ODE-Net, learning hamiltonian dynamics with Symplectic ODE-Net, and continuous normalizing flows (CNF).

****

TT-NF: Tensor Train Neural Fields

Anton Obukhov,Mikhail Usvyatsov,Christos Sakaridis,Konrad Schindler,Luc Van Gool

Learning neural fields has been an active topic in deep learning research, focusing, among other issues, on finding more compact and easy-to-fit representations. In this paper, we introduce a novel low-rank representation termed Tensor Train Neural Fields (TT-NF) for learning neural fields on dense regular grids and efficient methods for sampling from them. Our representation is a TT parameterization of the neural field, trained with backpropagation to minimize a non-convex objective. We analyze the effect of low-rank compression on the downstream task quality metrics in two settings. First, we demonstrate the efficiency of our method in a sandbox task of tensor denoising, which admits comparison with SVD-based schemes designed to minimize reconstruction error. Furthermore, we apply the pr oposed approach to Neural Radiance Fields, where the low-rank structure of the field corresponding to the best quality can be discovered only through learning.

****

Partial transportability for domain generalization

Alexis Bellot,Elias Bareinboim

Learning prediction models that generalize to related domains is one of the most fundamental challenges in artificial intelligence. There exists a growing literature that argues for learning invariant associations using data from multiple source domains. However, whether invariant predictors generalize to a given target domain depends crucially on the assumed structural changes between domains. Using the perspective of transportability theory, we show that invariance learning, and the settings in which invariant predictors are optimal in terms of worst-case losses, is a special case of a more general partial transportability task. Specifically, the partial transportability task seeks to identify / bound a conditional expectation $\mathbb E_{P_{\pi^*}}[y\mid\mathbf x]$ in an unseen domain π^* using knowledge of qualitative changes across domains in the form of causal graphs and data from source domains $\pi^1,\dots,\pi^k$. We show that solutions to this problem have a much wider generalization guarantee that subsumes those of invariance learning and other robust optimization methods that are inspired by causality. For computations in practice, we develop an algorithm that provably provides tight bounds asymptotically in the number of data samples from source domains for any partial transportability problem with discrete observables and illustrate its use on synthetic datasets.

****

Feint in Multi-Player Games

Junyu Liu,Wangkai Jin,Xiangjun Peng

This paper introduces the first formalization, implementation and quantitative evaluation of \feint in Multi-Player Games. Our work first formalizes \feint from the perspective of Multi-Player Games, in terms of the temporal, spatial and their collective impacts. The formalization is built upon \textit{Non-transitive Active Markov Game Model}, where \feint can have a considerable amount of impacts. Then, our work considers practical implementation details of \feint in Multi-Player Games, under the state-of-the-art progress of multi-agent modeling to date (namely Multi-Agent Reinforcement Learning). Finally, our work quantitatively examines the effectiveness of our design, and the results show that our design of Feint can (1) greatly improve the reward gains from the game; (2) significantly improve the diversity of Multi-Player Games; and (3) only incur negligible overheads in terms of time consumption. We conclude that our design of Feint is effective and practical, to make Multi-Player Games more interesting.

****

Succinct Compression: Lossless Compression for Fast and Memory-Efficient Deep Neural Network Inference

Yicun Duan,Xiangjun Peng

This paper introduces ``Succinct Compression”, a method to provide lossless compression of Deep Neural Network (DNN) models for fast and memory-efficient inference. The key insight of our method leverages the concept of \textit{Succinct Data Structures}, which supports fast queries without decompressing the compressed representations. Our method consists of three new insights. First, we introduce two basic building blocks to formulate DNN models, and how they can be extended to be synergistic with compressed models (e.g. pruned or quantized models). Then, we propose a scheme to enable mixed-formulation inference for different layers, to better extract its benefits. Finally, our method exploits a specialized execution pipeline to incorporate different model formulations for fast inference. We quantitatively demonstrate that: our method can (1) enable faster and more memory-efficient inference on uncompressed models; (2) be synergistic with a variety of structure-altered/unaltered compression schemes with better speedup and compression ratio, while preserving the accuracy; and (3) can outperform all other state-of-the-art Model Coding approaches.

****

BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection
Zehui Chen,Zhenyu Li,Shiquan Zhang,Liangji Fang,Qinhong Jiang,Feng Zhao

3D object detection from multiple image views is a fundamental and challenging task for visual scene understanding. Owing to its low cost and high efficiency, multi-view 3D object detection has demonstrated promising application prospects. However, accurately detecting objects through perspective views is extremely difficult due to the lack of depth information. Current approaches tend to adopt heavy backbones for image encoders, making them inapplicable for real-world deployment. Different from the images, LiDAR points are superior in providing spatial cues, resulting in highly precise localization. In this paper, we explore the incorporation of LiDAR-based detectors for multi-view 3D object detection. Instead of directly training a depth prediction network, we unify the image and LiDAR features in the Bird-Eye-View (BEV) space and adaptively transfer knowledge across non-homogenous representations in a teacher-student paradigm. To this end, we propose BEVDistill, a cross-modal BEV knowledge distillation (KD) framework for multi-view 3D object detection.

Extensive experiments demonstrate that the proposed method outperforms current KD approaches on a highly-competitive baseline, BEVFormer, without introducing any extra cost in the inference phase. Notably, our best model achieves 59.4 NDS on the nuScenes test leaderboard, achieving new state-of-the-arts in comparison with various image-based detectors. Code will be available at https://github.com/zehuichen123/BEVDistill.

****

Expanding Datasets With Guided Imagination
Yifan Zhang,Zhou Daquan,Bryan Hooi,Kai Wang,Jiashi Feng

The power of Deep Neural Networks (DNNs) depends heavily on the training data quantity, quality and diversity. However, in many real scenarios, it is costly and time-consuming to collect and annotate large-scale data. This has severely hindered the application of DNNs. To address this challenge, we explore a new task of dataset expansion, which seeks to automatically create new labeled samples to expand a small dataset. To this end, we present a Guided Imagination Framework (GIF) that leverages the recently developed big generative models (e.g., DALL-E2) to ``imagine'' and create informative new data from seed data to expand small datasets. Specifically, GIF conducts imagination by optimizing the latent features of seed data in a semantically meaningful space, which are fed into the generative models to generate photo-realistic images with new contents. For guiding the imagination towards creating samples useful for model training, we exploit the zero-shot recognition ability of CLIP and introduce three criteria to encourage informative sample generation, i.e., prediction consistency, entropy maximization and diversity promotion. With these essential criteria as guidance, GIF works well for expanding datasets in different domains, leading to 29.9\% accuracy gain on average over six natural image datasets, and 10.4\% accuracy gain on average over three medical image datasets. The source code will be made public.

****

ThinkSum: Probabilistic reasoning over sets using large language models
Batu Ozturkler,Nikolay Malkin,Zhen Wang,Nebojsa Jojic

Large language models (LLMs) have a substantial capacity for high-level analogical reasoning: reproducing patterns in linear text that occur in their training data (zero-shot evaluation) or in the provided context (few-shot in-context learning). However, recent studies show that even the largest LLMs fail in scenarios that require reasoning over multiple objects or facts or making sequences of logical deductions. We propose a two-stage probabilistic inference paradigm, ThinkSum, that reasons over sets of objects or facts in a structured manner. In the first stage (Think -- 'fast' retrieval of associations), a LLM is queried in parallel over a set of phrases extracted from the prompt or an auxiliary model call. In the second stage (Sum -- 'slow' probabilistic inference or reasoning), the results of these queries are aggregated to make the final prediction. We demonstrate the advantages of ThinkSum on the BIG-bench suite of evaluation tasks, achieving improvements over the state of the art using GPT-family models on ten difficult tasks, often with far smaller model variants. We compare and contrast ThinkS um with other proposed modifications to direct prompting of LLMs, such as variants of chain-of-thought prompting. We argue that because the probabilistic inference in ThinkSum is performed outside of calls to the LLM, ThinkSum is less sensitive to prompt design, yields more interpretable predictions, and can be flexibly combined with latent variable models to extract structured knowledge from LLMs.

****

Confidence and Dispersity Speak: Characterising Prediction Matrix for Unsupervised Accuracy Estimation

Weijian Deng,Yumin Suh,Stephen Gould,Liang Zheng

This work focuses on estimating how well a model performs on out-of-distribution (OOD) datasets without using labels. Our intuition is that a well-performing model should give predictions with high confidence and high dispersity. While recent methods study the prediction confidence, this work newly finds dispersity is another informative cue. Confidence reflects whether the individual prediction is certain; dispersity indicates how the overall predictions are distributed across all categories. To achieve a more accurate estimation, we propose to jointly consider these two properties by using the nuclear norm of the prediction matrix. In our experiments, we extensively validate the effectiveness of nuclear norm for various models (e.g., ViT and ConvNeXt), different datasets (e.g., ImageNet and CUB-200), and diverse types of distribution shifts (e.g., style shift and reproduction shift). We show that the nuclear norm is more accurate and robust in predicting OOD accuracy than existing methods. Lastly, we study the limitation of the nuclear norm and discuss potential directions.

****

Design of the topology for contrastive visual-textual alignment

Zhun Sun

Pre-training weakly related image-text pairs in the contrastive style shows great power in learning semantic aligning cross-modal models. The common choice to measure the distance between the feature representations of the image-text pairs is the cosine similarity, which can be considered as the negative inner product distance of features embedded on a sphere, mathematically. However, empirically, aligning image-text pairs on the spherical topology is vulnerable to the semantic ambiguity phenomenon resulting from the noise in the pre-training datasets. Specifically, under the noisy training data, instead of the optimal alignment-uniformity solution, the system would achieve an equilibrium (a gap between distances of positive and negative pairs), when the gradients for attraction and repulsion are neutralized. Although intuitively, the model should always find this equilibrium given a sufficiently long training scheme, its numerical values might be out of the distance range (e.g. [-1, 1] for the cosine similarity). In the practice of former studies, this problem is partly tackled by introducing a learnable softmax temperature parameter, in other words, by explicitly scaling the range of the distance function. In this work, we alternatively design the topology of embedding space and its endowed distance function. Motivated by studies that make use of Riemannian geometry for visual tasks, we propose a rather simple solution to address the aforementioned equilibrium problem. That is, we map the feature representations onto the oblique manifold endowed with the negative inner product as the distance function. In the experimental analysis, we show that we can improve the baseline performance by a large margin (e.g. 4\% in the zero-shot image to text retrieval task) by changing only two lines of the training codes.

****

Slimmable Networks for Contrastive Self-supervised Learning

Shuai Zhao,Xiaohan Wang,Linchao Zhu,Yi Yang

Self-supervised learning makes great progress in large model pre-training but suffers in training small models. Previous solutions to this problem mainly rely on knowledge distillation and indeed have a two-stage learning procedure: first train a large teacher model, then distill it to improve the generalization ability of small ones. In this work, we present a new one-stage solution to obtain pre-trained small models without extra teachers: slimmable networks for contrastive self-supervised learning (SlimCLR). A slimmable network contains a full network and several weight-sharing sub-networks. We can pre-train for only one time and obtain various networks including small ones with low computation costs. However, in self-supervised cases, the interference between weight-sharing networks leads to severe performance degradation. One evidence of the interference is gradient imbalance: a small proportion of parameters produces dominant gradients during backpropagation, and the main parameters may not be fully optimized. The divergence in gradient directions of various networks may also cause interference between networks. To overcome these problems, we make the main parameters produce dominant gradients and provide consistent guidance for sub-networks via three techniques: slow start training of sub-networks, online distillation, and loss re-weighting according to model sizes. Besides, a switchable linear probe layer is applied during linear evaluation to avoid the interference of weight-sharing linear layers. We instantiate SlimCLR with typical contrastive learning frameworks and achieve better performance than previous arts with fewer parameters and FLOPs.

****

A Multi-Grained Self-Interpretable Symbolic-Neural Model For Single/Multi-Labeled Text Classification

Xiang Hu,XinYu KONG,Kewei Tu

Deep neural networks based on layer-stacking architectures have historically suffered from poor inherent interpretability. Meanwhile, symbolic probabilistic models function with clear interpretability, but how to combine them with neural networks to enhance their performance remains to be explored. In this paper, we try to marry these two systems for text classification via a structured language model. We propose a Symbolic-Neural model that can learn to explicitly predict class labels of text spans from a constituency tree without requiring any access to span-level gold labels. As the structured language model learns to predict constituency trees in a self-supervised manner, only raw texts and sentence-level labels are required as training data, which makes it essentially a general constituent-level self-interpretable classification model. Our experiments demonstrate that our approach could achieve good prediction accuracy in downstream tasks. Meanwhile, the predicted span labels are consistent with human rationales to a certain degree.

****

Suppressing the Heterogeneity: A Strong Feature Extractor for Few-shot Segmentation

Zhengdong Hu,Yifan Sun,Yi Yang

This paper tackles the Few-shot Semantic Segmentation (FSS) task with focus on learning the feature extractor. Somehow the feature extractor has been overlooked by recent state-of-the-art methods, which directly use a deep model pretrained on ImageNet for feature extraction (without further fine-tuning). Under this background, we think the FSS feature extractor deserves exploration and observe the heterogeneity (i.e., the intra-class diversity in the raw images) as a critical challenge hindering the intra-class feature compactness. The heterogeneity has three levels from coarse to fine: 1) Sample-level: the inevitable distribution gap between the support and query images makes them heterogeneous from each other. 2) Region-level: the background in FSS actually contains multiple regions with different semantics. 3) Patch-level: some neighboring patches belonging to a same class may appear quite different from each other. Motivated by these observations, we propose a feature extractor with Multi-level Heterogeneity Suppressing (MuHS). MuHS leverages the attention mechanism in transformer backbone to effectively suppress all these three-level heterogeneity. Concretely, MuHS reinforces the attention / interaction between different samples (query and support), different regions and neighboring patches by constructing cross-sample attention, cross-region interaction and a novel masked image segmentation (inspired by the recent masked image modeling), respectively. We empirically show that 1) MuHS brings consistent improvement for various FSS heads and 2) using a simple linear classification head, MuHS sets new states of the art on multiple FSS datasets, validating the importance of FSS feature learning.

****

Communication Efficient Fair Federated Recommender System

KIRANDEEP KAUR,Sujit Gujar,Shweta Jain

Federated Recommender Systems (FRSs) aim to provide recommendations to clients in a distributed manner with privacy preservation. FRSs suffer from high communication costs due to the communication between the server and many clients. Some past literature on federated supervised learning shows that sampling clients randomly improve communication efficiency without jeopardizing accuracy. However, each user is considered a separate client in FRS and clients communicate only item gradients. Thus, incorporating random sampling and determining the number of clients to be sampled in each communication round to retain the model's accuracy in FRS becomes challenging. This paper provides sample complexity bounds on the number of clients that must be sampled in an FRS to preserve accuracy. Next, we consider the issue of demographic bias in FRS, quantified as the difference in the average error rates across different groups. Supervised learning algorithms mitigate the group bias by adding the fairness constraint in the training loss, which requires sharing protected attributes with the server. This is prohibited in a federated setting to ensure clients' privacy. We design \ouralgo, a Random Sampling based Fair Federated Recommender System, which trains to achieve a fair global model. In addition, it also trains local clients towards a fair global model to reduce demographic bias at the client level without the need to share their protected attributes. We empirically demonstrate all our results across the two most popular real-world datasets (ML1M, ML100k) and different sensitive features (age and gender) to prove that RS-FairFRS helps reduce communication cost and demographic bias with improved model accuracy.

****

Grassmannian Class Representation in Deep Learning

Haoqi Wang,Zhizhong Li,Wayne Zhang

We generalize the class representative vector found in deep classification networks to linear subspaces and show that the new formulation enables the simultaneous enhancement of the inter-class discrimination and intra-class feature variation. Traditionally, the logit is computed by the inner product between a feature and the class vector. In our modeling, classes are subspaces and the logit is defined as the norm of the projection from a feature onto the subspace. Since the set of subspaces forms Grassmann manifolds, finding the optimal subspace representation for classes is to optimize the loss on a Grassmannian. We integrate the Riemannian SGD into existing deep learning frameworks such that the class subspaces in a Grassmannian are jointly optimized with other model parameters in Euclidean. Compared to the vector form, subspaces have two appealing properties: they can be multi-dimensional and they are scaleless. Empirically, we reveal that these distinct characteristics improve various tasks. (1) Image classification. The new formulation brings the top-1 accuracy of ResNet50-D on ImageNet-1K from 78.04% to 79.37% using the standard augmentation in 100 training epochs. This confirms that the representative capability of subspaces is more powerful than vectors. (2) Feature transfer. Subspaces provide freedom for features to vary and we observed that the intra-class variability of features increases when the subspace dimensions are larger. Consequently, the quality of features is better for downstream tasks. The average transfer accuracy across 6 datasets improves from 77.98% to 80.12% compared to the strong baseline of vanilla softmax. (3) Long-tail classification. The scaleless property of subspaces benefits classification in the long-tail scenario and improves the accuracy of ImageNet-LT from 46.83% to 48.94% compared to the standard formulation. With these encouraging results, we believe that more applications could benefit from the Grassmannian class representation. Codes will be released.

****

Refining Visual Representation for Generalized Zero-Shot Recognition through Implicit-Semantics-Guided Metric Learning

YanHe Chen,Mei-Chen Yeh

Deep metric learning (DML) is effective to address the large intra- and the smal l inter-class variation problem in visual recognition; however, when applied for generalized zero-shot learning (GZSL) in which the label of a target image may belong to an unseen category, this technique can be easily biased towards seen classes. Alternatively in GZSL some form of semantic space is available, which plays an important role in relating seen and unseen classes and is widely used to guide the learning of visual representation. To take advantage of DML while avoiding overfitting to seen classes, we propose a novel representation learning framework$\textemdash$Metric Learning with Implicit Semantics (MLIS)$\textemdash$to refine discriminative and generalizable visual features for GZSL. Specifically, we disentangle the effects of semantics on feature extractor and image classification of the model, so that semantics only participate in feature learning, and classification only uses the refined visual features. We further relax the visual-semantic alignment requirement, avoiding performing pair-wise comparisons between the image and the class embeddings. Experimental results demonstrate that the proposed MLIS framework bridges DML and GZSL. It achieves state-of-the-art performance, and is robust and flexible to the integration with several metric learning based loss functions.

****

Reward Learning with Trees: Methods and Evaluation

Tom Bewley,Jonathan Lawry,Arthur Richards,Rachel Craddock,Ian Henderson

Recent efforts to learn reward functions from human feedback have tended to use deep neural networks, whose lack of transparency hampers our ability to explain agent behaviour or verify alignment. We explore the merits of learning intrinsically interpretable tree models instead. We develop a recently proposed method for learning reward trees from preference labels, and show it to be broadly competitive with neural networks on challenging high-dimensional tasks, with good robustness to limited or corrupted data. Having found that reward tree learning can be done effectively in complex settings, we then consider why it should be used, demonstrating that the interpretable reward structure gives significant scope for traceability, verification and explanation.

****

Achieve the Minimum Width of Neural Networks for Universal Approximation

Yongqiang Cai

The universal approximation property (UAP) of neural networks is fundamental for deep learning, and it is well known that wide neural networks are universal approximators of continuous functions within both the L^p norm and the continuous/uniform norm. However, the exact minimum width, $w_{\min}$, for the UAP has not been studied thoroughly. Recently, using a decoder-memorizer-encoder scheme, \citet{Park2021Minimum} found that $w_{\min} = \max(d_x+1,d_y)$ for both the L^p-UAP of ReLU networks and the C-UAP of ReLU+STEP networks, where d_x,d_y are the input and output dimensions, respectively. In this paper, we consider neural networks with an arbitrary set of activation functions. We prove that both C-UAP and L^p-UAP for functions on compact domains share a universal lower bound of the minimal width; that is, $w^*_{\min} = \max(d_x,d_y)$. In particular, the critical width, $w^*_{\min}$, for L^p-UAP can be achieved by leaky-ReLU networks, provided that the input or output dimension is larger than one. Our construction is based on the approximation power of neural ordinary differential equations and the ability to approximate flow maps by neural networks. The nonmonotone or discontinuous activation functions case and the one-dimensional case are also discussed.

****

H2RBox: Horizontal Box Annotation is All You Need for Oriented Object Detection

Xue Yang,Gefan Zhang,Wentong Li,Yue Zhou,Xuehui Wang,Junchi Yan

Oriented object detection emerges in many applications from aerial images to autonomous driving, while many existing detection benchmarks are annotated with horizontal bounding box only which is also less costive than fine-grained rotated box, leading to a gap between the readily available training corpus and the rising demand for oriented object detection. This paper proposes a simple yet effective oriented object detection approach called H2RBox merely using horizontal box annotation for weakly-supervised training, which closes the above gap and shows competitive performance even against those trained with rotated boxes. The cores of our method are weakly- and self-supervised learning, which predicts the angle of the object by learning the consistency of two different views. To our best knowledge, H2RBox is the first horizontal box annotation-based oriented object detector. Compared to an alternative i.e. horizontal box-supervised instance segmentation with our post adaption to oriented object detection, our approach is not susceptible to the prediction quality of mask and can perform more robustly in complex scenes containing a large number of dense objects and outliers. Experimental results show that H2RBox has significant performance and speed advantages over horizontal box-supervised instance segmentation methods, as well as lower memory requirements. While compared to rotated box-supervised oriented object detectors, our method shows very close performance and speed. The source code is available at PyTorch-based \href{https://github.com/yangxue0827/h2rbox-mmrotate}{MMRotate} and Jittor-based \href{https://github.com/yangxue0827/h2rbox-jittor}{JDet}.

****

Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling

Keyu Tian,Yi Jiang,qishuai diao,Chen Lin,Liwei Wang,Zehuan Yuan

We identify and overcome two key obstacles in extending the success of BERT-style pre-training, or masked image modeling, to convolutional networks (convnets): (i) convolution operation cannot handle irregular, randomly masked input images; (ii) the single-scale nature of BERT pre-training is inconsistent with convnet’s hierarchical structure. For (i), we treat unmasked pixels as sparse voxels of 3D point clouds and use sparse convolution to encode. This is the first use of sparse convolution for 2D masked modeling. For (ii), we develop a hierarchical decoder to reconstruct images from multi-scale encoded features. Our method, called Sparse masKed modeling (SparK), is general: it can be used directly on any convolutional model without backbone modifications. We validate it on both classical (ResNet) and modern (ConvNeXt) models: on three downstream tasks, it surpasses both state-of-the-art contrastive learning and transformer-based masked modeling by similarly large margins (around +1.0%). The improvements on object detection and instance segmentation are more significant (up to +3.5%), validating the strong transferability of features learned. We also find SparK’s favorable scaling behavior by observing more gains on larger networks. All of these findings support the promising future of generative pre-training on convnets. Both codes and pre-trained models have been released at https://github.com/keyu-tian/SparK.

****

Functional Relation Field: A Model-Agnostic Framework for Multivariate Time Series Forecasting

Bing Yu,Ting Li,Jianguo Li,Bin Dong,Zhanxing Zhu

In multivariate time series forecasting, the most popular strategy for modeling the relationship between multiple time series is the construction of graph, where each time series is represented as a node and related nodes are connected by edges, i.e. spatial-temporal graph neural networks. The graph structure is either given apriori or learned based the similarity between nodes. However, the relationship between multiple time series is typically complicated, for instance, the sum of outflows from upstream nodes may be equal to the inflows of downstream nodes. Such relations widely exist in many real-world multivariate time series forecasting scenarios, yet are far from well studied. In these cases, graph might only be a crude description on the dependency between nodes. To this end, we explore a new framework to model the inter-node relationship in a more precise way based our proposed inductive bias for graphs, Functional Relation Field, where a group of functions parameterized by neural networks are learned to characterize the dependency between multiple time series. These learned functions are versatile: they can then be used to discover the underlying graph structure by identifying the most relevant neighbors of the target node; and on the other hand, the learned functions will form a “field” where the nodes in the backbone prediction networks are enforced to satisfy the constraints defined by these functions. Th e experiment is conducted on one toy dataset to show our approach can well recover the true constraint relationship between nodes. And two real-world MiniApp calling traffic and road network datasets are also considered with various different backbone networks. Results show that the prediction error can be reduced remarkably with the aid of the proposed functional relation field framework.

****

Motion-inductive Self-supervised Object Discovery in Videos

Shuangrui Ding,Weidi Xie,Yabo Chen,Rui Qian,XIAOPENG ZHANG,Hongkai Xiong,Qi Tian

In this paper, we consider the task of unsupervised object discovery in videos. Previous works have shown promising results via processing optical flows to segment objects. However, taking flow as input brings about two drawbacks. First, flow cannot capture sufficient cues when objects remain static or partially occluded. Second, it is challenging to establish temporal coherency from flow-only input, due to the missing texture information. To tackle these limitations, we propose a model for directly processing consecutive RGB frames, and infer the optical flow between any pair of frames using a layered representation, with the opacity channels being treated as the segmentation. Additionally, to enforce object permanence, we apply temporal consistency loss on the inferred masks from randomly-paired frames, which refer to the motions at different paces, and encourage the model to segment the objects even if they may not move at the current time point. Experimentally, we demonstrate superior performance over previous state-of-the-art methods on three public video segmentation datasets (DAVIS2016, SegTrackv2, and FBMS-59), while being computationally efficient by avoiding the overhead of computing optical flow as input.

****

Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment

Qiang Chen,Xiaokang Chen,Jian Wang,Shan Zhang,Haocheng Feng,Junyu Han,Errui Ding,Gang Zeng,Jingdong Wang

Detection Transformer (DETR) relies on one-to-one assignment for end-to-end object detection and lacks the capability of exploiting multiple positive object queries. We present a novel DETR training approach, named {\em Group DETR}, to support one-to-many assignment in a group-wise manner. To achieve it, we make simple modifications during training: (i) adopt K groups of object queries; (ii) conduct decoder self-attention on each group of object queries with the same parameters; (iii) perform one-to-one assignment for each group, leading to K positive object queries for each ground-truth object. In inference, we only use one group of object queries, making no modifications to model architectures and inference processes. Group DETR is a versatile training method and is applicable to various DETR variants. Our experiments show that Group DETR significantly speeds up the training convergences and improves the performances of various DETR-based methods.

****

Transcendental Idealism of Planner: Evaluating Perception from Planning Perspective for Autonomous Driving

Weixin Li,Xiaodong Yang

Evaluating the performance of perception module in autonomous driving is one of the most critical tasks in developing these complex intelligent systems. While module-level unit test methodologies adopted from traditional computer vision tasks are viable to a certain extent, it still remains far less explored to evaluate how changes in a perception module can impact the planning of an autonomous vehicle in a consistent and holistic manner. In this work, we propose a principled framework that provides a coherent and systematic understanding of how perception modules affect the planning of an autonomous vehicle that actually controls the vehicle. Specifically, planning of an autonomous vehicle is formulated as an expected utility maximisation problem, where all input signals from upstream modules jointly provide a world state description, and the planner aims to find the optimal action to execute by finding the solution to maximise the expected utility determined by both the world state and the action. We show that, under some mild conditions, the objective function can be represented as an inner product between the world state description and the utility function in a Hilbert space.

This geometric interpretation enables a novel way to formulate, analyse and evaluate the impact of noise in world state estimation on the solution to the problem, and leads to a universal quantitative metric for such purpose. The whole framework resembles the idea of transcendental idealism in the classical philosophy literature, which gives the name to our approach.
****
Pushing the Limits of Fewshot Anomaly Detection in Industry Vision: Graphcore
Guoyang Xie,Jinbao Wang,Jiaqi Liu,Yaochu Jin,Feng Zheng
In the area of few-shot anomaly detection (FSAD), efficient visual feature plays an essential role in the memory bank $\mathcal{M}$-based methods. However, these methods do not account for the relationship between the visual feature and its rotated visual feature, drastically limiting the anomaly detection performance. To push the limits, we reveal that rotation-invariant feature property has a significant impact on industrial-based FSAD. Specifically, we utilize graph representation in FSAD and provide a novel visual isometric invariant feature (VIIF) as an anomaly measurement feature. As a result, VIIF can robustly improve the anomaly discriminating ability and can further reduce the size of redundant features stored in $\mathcal{M}$ by a large amount. Besides, we provide a novel model GraphCore via VIIFs that can fast implement unsupervised FSAD training and improve the performance of anomaly detection. A comprehensive evaluation is provided for comparing GraphCore and other SOTA anomaly detection models under our proposed few-shot anomaly detection setting, which shows GraphCore can increase average AUC by 5.8%, 4.1%, 3.4%, and 1.6% on MVTec AD and by 25.5%, 22.0%, 16.9%, and 14.1% on MPDD for 1, 2, 4, and 8-shot cases, respectively.
****
Evaluating Weakly Supervised Object Localization Methods Right? A Study on Heatmap-based XAI and Neural Backed Decision Tree
Erico Tjoa,Cuntai Guan
Choe et al have investigated several aspects of Weakly Supervised Object Localization (WSOL) with only image label. They addressed the ill-posed nature of the problem and showed that WSOL has not significantly improved beyond the baseline method class activation mapping (CAM). We report the results of similar experiments on ResNet50 with some crucial differences: (1) we perform WSOL using heatmap-based eXplanaible AI (XAI) methods (2) our model is not class agnostic since we are interested in the XAI aspect as well. Under similar protocol, we find that XAI methods perform WSOL with very sub-standard MaxBoxAcc scores. The experiment is then repeated for the same model trained with Neural Backed Decision Tree (NBDT) and we found that vanilla CAM yields significantly better WSOL performance after NBDT training.
****
Representation Learning for Low-rank General-sum Markov Games
Chengzhuo Ni,Yuda Song,Xuezhou Zhang,Zihan Ding,Chi Jin,Mengdi Wang
We study multi-agent general-sum Markov games with nonlinear function approximation. We focus on low-rank Markov games whose transition matrix admits a hidden low-rank structure on top of an unknown non-linear representation. The goal is to design an algorithm that (1) finds an ε-equilibrium policy sample efficiently without prior knowledge of the environment or the representation, and (2) permits a deep-learning friendly implementation. We leverage representation learning and present a model-based and a model-free approach to construct an effective representation from collected data. For both approaches, the algorithm achieves a sample complexity of poly$(H,d,A,1/\varepsilon)$, where H is the game horizon, d is the dimension of the feature vector, A is the size of the joint action space and ε is the optimality gap. When the number of players is large, the above sample complexity can scale exponentially with the number of players in the worst case. To address this challenge, we consider Markov Games with a factorized transition structure and present an algorithm that escapes such exponential scaling. To our best knowledge, this is the first sample-efficient algorithm for multi-agent general-sum Markov games that incorporates (non-linear) function approximation. We accompany our theoretical result with a neural network-based implementation of our algorithm and evaluate it against the wide

ly used deep RL baseline, DQN with fictitious play.

****

Multi-Domain Long-Tailed Learning by Augmenting Disentangled Representations

Huaxiu Yao,Xinyu Yang,Allan Zhou,Chelsea Finn

There is an inescapable long-tailed class-imbalance issue in many real-world classification problems. Existing long-tailed classification methods focus on the single-domain setting, where all examples are drawn from the same distribution. However, real-world scenarios often involve multiple domains with distinct imbalanced class distributions. We study this multi-domain long-tailed learning problem and aim to produce a model that generalizes well across all classes and domains. Towards that goal, we introduce TALLY, which produces invariant predictors by balanced augmenting hidden representations over domains and classes. Built upon a proposed selective balanced sampling strategy, TALLY achieves this by mixing the semantic representation of one example with the domain-associated nuisances of another, producing a new representation for use as data augmentation. To improve the disentanglement of semantic representations, TALLY further utilizes a domain-invariant class prototype that averages out domain-specific effects. We evaluate TALLY on four long-tailed variants of classical domain generalization benchmarks and two real-world imbalanced multi-domain datasets. The results indicate that TALLY consistently outperforms other state-of-the-art methods in both subpopulation shift and domain shift.

****

Surgical Fine-Tuning Improves Adaptation to Distribution Shifts

Yoonho Lee,Annie S Chen,Fahim Tajwar,Ananya Kumar,Huaxiu Yao,Percy Liang,Chelsea Finn

A common approach to transfer learning under distribution shift is to fine-tune the last few layers of a pre-trained model, preserving learned features while also adapting to the new task. This paper shows that in such settings, selectively fine-tuning a subset of layers (which we term surgical fine-tuning) matches or outperforms commonly used fine-tuning approaches. Moreover, the type of distribution shift influences which subset is more effective to tune: for example, for image corruptions, fine-tuning only the first few layers works best. We validate our findings systematically across seven real-world data tasks spanning three types of distribution shifts. Theoretically, we prove that for two-layer neural networks in an idealized setting, first-layer tuning can outperform fine-tuning all layers. Intuitively, fine-tuning more parameters on a small target dataset can cause information learned during pre-training to be forgotten, and the relevant information depends on the type of shift.

****

Diversify and Disambiguate: Out-of-Distribution Robustness via Disagreement

Yoonho Lee,Huaxiu Yao,Chelsea Finn

Real-world machine learning problems often exhibit shifts between the source and target distributions, in which source data does not fully convey the desired behavior on target inputs. Different functions that achieve near-perfect source accuracy can make differing predictions on test inputs, and such ambiguity makes robustness to distribution shifts challenging. We propose DivDis, a simple two-stage framework for identifying and resolving ambiguity in data. DivDis first learns a diverse set of hypotheses that achieve low source loss but make differing predictions on target inputs. We then disambiguate by selecting one of the discovered functions using additional information, for example, a small number of target labels. Our experimental evaluation shows improved performance in subpopulation shift and domain generalization settings, demonstrating that DivDis can scalably adapt to distribution shifts in image and text classification benchmarks.

****

MaSS: Multi-attribute Selective Suppression

Chun-Fu Chen,Shaohan Hu,Zhonghao Shi,Prateek Gulati,Bill Moriarty,Marco Pistoia,Vincenzo Piuri,Pierangela Samarati

The recent rapid advances in the development and deployment of machine learning technologies largely depend on the vast richness of data available today, in terms of both the quantity and the rich content contained within. For example, biom

etric data such as images and voices could reveal people's attributes like age, gender, sentiment, and origin, whereas location/motion data could be used to infer people's activity levels, transportation modes, and life habits. Along with the new services and applications enabled by such technological advances, various governmental policies are put in place to regulate such data usage and protect people's privacy and rights. As a result, data owners often opt for simple data obfuscation (e.g., blur people's faces in images) or withholding data altogether, which leads to severe data quality degradation and greatly limits the data's potential utility.
Aiming for a sophisticated mechanism which gives data owners fine-grained control while retaining the maximal degree of data utility, we propose Multi-attribute Selective Suppression, or MaSS, a general framework for performing precisely targeted data surgery to simultaneously suppress any selected set of attributes while preserving the rest for downstream machine learning tasks. MaSS learns a data modifier through adversarial games between two sets of networks, where one is aimed at suppressing selected attributes, and the other ensures the retention of the rest of the attributes via general contrastive loss as well as explicit classification metrics. We carried out an extensive evaluation of our proposed method using multiple datasets from different domains including facial images, voice audio, and video clips, and obtained highly promising results in MaSS' generalizability and capability of drastically suppressing targeted attributes (e.g., reducing inference on such attributes to random guess) while imposing virtually no impact on the data's usability in other downstream ML tasks.
****

Neural Attention Memory

Hyoungwook Nam,Seung Byum Seo

Scaled dot-product attention has become the essence of state-of-the-art deep neural networks for various machine learning tasks. Though its ubiquitous accomplishments, it is inefficient for long sequence tasks and problematic for tasks requiring memory states such as compositional generalization. We propose a novel perspective of the attention mechanism by reinventing it as a memory architecture for neural networks, namely Neural Attention Memory (NAM). NAM follows the same query-key-value structure by constructing a memory matrix while reducing its computational complexity from quadratic to linear to the sequence length. NAM writes a memory matrix via the sum of outer products of value and unit key vectors, and reads it by multiplying the matrix with a unit query vector. Indeed, we show that our normalized outer-product attention mechanism is mathematically equivalent to the conventional attention mechanism. Then, we evaluate a NAM-based Transformer on long-range arena tasks and demonstrate its efficiency and efficacy. Finally, we propose two NAM-based memory-augmented neural networks, namely Long Short-Term Attention Memory (LSAM) and NAM Turing Machine (NAM-TM), and test their compositional generalization capability using four different tasks. LSAM replaces LSTM's long-term cell state with NAM memory matrix and NAM-TM implements a Turing tape data structure using NAM read/write primitives. The experimental results show that the proposed models outperform traditional Transformer and LSTM, as well as DNC. NAM opens up possibilities in diverse machine learning research problems, including hierarchical data modeling, efficient edge inference, and few-shot learning.
****

Meta Optimal Transport

Brandon Amos,Samuel Cohen,Giulia Luise,Ievgen Redko

We study the use of amortized optimization to predict optimal transport (OT) maps from the input measures, which we call Meta OT. This helps repeatedly solve similar OT problems between different measures by leveraging the knowledge and information present from past problems to rapidly predict and solve new problems. Otherwise, standard methods ignore the knowledge of the past solutions and suboptimally re-solve each problem from scratch. We instantiate Meta OT models in discrete and continuous (Wasserstein-2) settings between images, spherical data, and color palettes and use them to improve the computational time of standard OT solvers by multiple orders of magnitude.

****

On amortizing convex conjugates for optimal transport

Brandon Amos

This paper focuses on computing the convex conjugate operation that arises when solving Euclidean Wasserstein-2 optimal transport problems. This conjugation, which is also referred to as the Legendre-Fenchel conjugate or c-transform,is considered difficult to compute and in practice,Wasserstein-2 methods are limited by not being able to exactly conjugate the dual potentials in continuous space. To overcome this, the computation of the conjugate can be approximated with amortized optimization, which learns a model to predict the conjugate. I show that combining amortized approximations to the conjugate with a solver for fine-tuning significantly improves the quality of transport maps learned for the Wasserstein-2 benchmark by Korotin et al. (2021a) and is able to model many 2-dimensional couplings and flows considered in the literature. All of the baselines, methods, and solvers in this paper are available at http://github.com/facebookresearch/w2ot.

****

Example-based Planning via Dual Gradient Fields

Mingdong Wu,fangwei zhong,Yizhou Wang,Hao Dong

Path planning is one of the key abilities of an intelligent agent. However, both the learning-based and sample-based planners remain to require explicitly defining the task by manually designing the reward function or optimisation objectives, which limits the scope of implementation. Formulating the path planning problem from a new perspective, Example-based planning is to find the most efficient path to increase the likelihood of the target distribution by giving a set of target examples. In this work, we introduce Dual Gradient Fields (DualGFs), an offline-learning example-based planning framework built upon score matching. There are two gradient fields in DualGFs: a target gradient field that guides task completion and a support gradient field that ensures moving with environmental constraints. In the learning process, instead of interacting with the environment, the agents are trained with two offline examples, i.e., the target gradients and support gradients are trained by target examples and support examples, respectively. The support examples are randomly sampled from free space, e.g., states without collisions. DualGF is a weighted mixture of the two fields, combining the merits of the two fields together. To update the mixing ratio adaptively, we further propose a fields-balancing mechanism based on Lagrangian-Relaxation. Experimental results across four tasks (navigation, tracking, particle rearrangement, and room rearrangement) demonstrate the scalability and effectiveness of our method.

****

DualAfford: Learning Collaborative Visual Affordance for Dual-gripper Manipulation

Yan Zhao,Ruihai Wu,Zhehuan Chen,Yourong Zhang,Qingnan Fan,Kaichun Mo,Hao Dong

It is essential yet challenging for future home-assistant robots to understand and manipulate diverse 3D objects in daily human environments. Towards building scalable systems that can perform diverse manipulation tasks over various 3D shapes, recent works have advocated and demonstrated promising results learning visual actionable affordance, which labels every point over the input 3D geometry with an action likelihood of accomplishing the downstream task (e.g., pushing or picking-up). However, these works only studied single-gripper manipulation tasks, yet many real-world tasks require two hands to achieve collaboratively. In this work, we propose a novel learning framework, DualAfford, to learn collaborative affordance for dual-gripper manipulation tasks. The core design of the approach is to reduce the quadratic problem for two grippers into two disentangled yet interconnected subtasks for efficient learning. Using the large-scale PartNet-Mobility and ShapeNet datasets, we set up four benchmark tasks for dual-gripper manipulation. Experiments prove the effectiveness and superiority of our method over three baselines. We will release code and data upon acceptance.

****

GraphCG: Unsupervised Discovery of Steerable Factors in Graphs

Shengchao Liu,Chengpeng Wang,Weili Nie,Hanchen Wang,Jiarui Lu,Bolei Zhou,Jian Tang

Deep generative models have been widely developed for graph data such as molecular graphs and point clouds. Yet, much less investigation has been carried out on understanding the learned latent space of deep graph generative models. Such understandings can open up a unified perspective and provide guidelines for essential tasks like controllable generation. To this end, this work develops a method called GraphCG for unsupervised discovery of steerable factors in latent space of deep graph generative models. We first examine the representation space of the recent deep generative models trained for graph data, and observe that the learned representation space is not perfectly disentangled. Thus, our method is designed for discovering steerable factors of graph data in a model-agnostic and task-agnostic manner. Specifically, GraphCG learns the semantic-rich directions via maximizing the corresponding mutual information, where the edited graph along the same direction will possess certain steerable factors. We conduct experiments on two types of graph data, molecular graphs and point clouds. Both the quantitative and qualitative results show the effectiveness of GraphCG for discovering steerable factors.

****

Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching

Shengchao Liu,Hongyu Guo,Jian Tang

Molecular representation pretraining is critical in various applications for drug and material discovery due to the limited number of labeled molecules, and most existing work focuses on pretraining on 2D molecular graphs. However, the power of pretraining on 3D geometric structures has been less explored. This is owing to the difficulty of finding a sufficient proxy task that can empower the pretraining to effectively extract essential features from the geometric structures. Motivated by the dynamic nature of 3D molecules, where the continuous motion of a molecule in the 3D Euclidean space forms a smooth potential energy surface, we propose GeoSSL, a 3D coordinate denoising pretraining framework to model such an energy landscape. Further by leveraging an SE(3)-invariant score matching method, we propose GeoSSL-DDM in which the coordinate denoising proxy task is effectively boiled down to denoising the pairwise atomic distances in a molecule. Our comprehensive experiments confirm the effectiveness and robustness of our proposed method.

****

SIMPLE: Specialized Model-Sample Matching for Domain Generalization

Ziyue Li,Kan Ren,XINYANG JIANG,Yifei Shen,Haipeng Zhang,Dongsheng Li

In domain generalization (DG), most existing methods aspire to fine-tune a specific pretrained model through novel DG algorithms. In this paper, we propose an alternative direction, i.e., to efficiently leverage a pool of pretrained models without fine-tuning. Through extensive empirical and theoretical evidence, we demonstrate that (1) pretrained models have possessed generalization to some extent while there is no single best pretrained model across all distribution shifts, and (2) out-of-distribution (OOD) generalization error depends on the fitness between the pretrained model and unseen test distributions. This analysis motivates us to incorporate diverse pretrained models and to dispatch the best matched models for each OOD sample by means of recommendation techniques. To this end, we propose SIMPLE, a specialized model-sample matching method for domain generalization. First, the predictions of pretrained models are adapted to the target domain by a linear label space transformation. A matching network aware of model specialty is then proposed to dynamically recommend proper pretrained models to predict each test sample. The experiments on DomainBed show that our method achieves significant performance improvements (up to 12.2% for individual dataset and 3.9% on average) compared to state-of-the-art (SOTA) methods and further achieves 6.1% gain via enlarging the pretrained model pool. Moreover, our method is highly efficient and achieves more than 1000 times training speedup compared to the conventional DG methods with fine-tuning a pretrained model. Code and supplemental materials are available at https://seqml.github.io/simple.

****

The Augmented Image Prior: Distilling 1000 Classes by Extrapolating from a Single Image
Yuki M Asano,Aaqib Saeed
What can neural networks learn about the visual world when provided with only a single image as input? While any image obviously cannot contain the multitudes of all existing objects, scenes and lighting conditions -- within the space of all $256^{3\cdot224\cdot224}$ possible 224-sized square images, it might still provide a strong prior for natural images. To analyze this ``augmented image prior'' hypothesis, we develop a simple framework for training neural networks from scratch using a single image and augmentations using knowledge distillation from a supervised pretrained teacher. With this, we find the answer to the above question to be: `surprisingly, a lot'. In quantitative terms, we find accuracies of 94%/74% on CIFAR-10/100, 69\% on ImageNet, and by extending this method to video and audio, 51% on Kinetics-400 and 84\% on SpeechCommands. In extensive analyses spanning 13 datasets, we disentangle the effect of augmentations, choice of data and network architectures and also provide qualitative evaluations that include lucid ``panda neurons'' in networks that have never even seen one.
****
Protein structure generation via folding diffusion
Kevin Eric Wu,Kevin K Yang,Rianne van den Berg,James Zou,Alex Xijie Lu,Ava P Amini
The ability to computationally generate novel yet physically foldable protein structures could lead to new biological discoveries and new treatments targeting yet incurable diseases. Despite recent advances in protein structure prediction, directly generating diverse, novel protein structures from neural networks remains difficult. In this work, we present a new diffusion-based generative model that designs protein backbone structures via a procedure that mirrors the native folding process. We describe protein backbone structure as a series of consecutive angles capturing the relative orientation of the constituent amino acid residues, and generate new structures by denoising from a random, unfolded state towards a stable folded structure. Not only does this mirror how proteins biologically twist into energetically favorable conformations, the inherent shift and rotational invariance of this representation crucially alleviates the need for complex equivariant networks. We train a denoising diffusion probabilistic model with a simple transformer backbone and demonstrate that our resulting model unconditionally generates highly realistic protein structures with complexity and structural patterns akin to those of naturally-occurring proteins. As a useful resource, we release the first open-source codebase and trained models for protein structure diffusion.
****
Delving into Semantic Scale Imbalance
Yanbiao Ma,Licheng Jiao,Fang Liu,Yuxin Li,Shuyuan Yang,Xu Liu
Model bias triggered by long-tailed data has been widely studied. However, measure based on the number of samples cannot explicate three phenomena simultaneously: (1) Given enough data, the classification performance gain is marginal with additional samples. (2) Classification performance decays precipitously as the number of training samples decreases when there is insufficient data. (3) Model trained on sample-balanced datasets still has different biases for different classes. In this work, we define and quantify the semantic scale of classes, which is equivalent to the feature diversity of classes. It is exciting to find experimentally that there is a marginal effect of semantic scale, which perfectly describes the first two phenomena. Further, the quantitative measurement of semantic scale imbalance is proposed, which can accurately reflect model bias on multiple datasets, even on sample-balanced data, revealing a novel perspective for the study of class imbalance. Due to the prevalence of semantic scale imbalance, we propose semantic-scale-balanced learning, including a general loss improvement scheme and a dynamic re-weighting training framework that overcomes the challenge of calculating semantic scales in real-time during iterations. Comprehensive experiments show that dynamic semantic-scale-balanced learning consistently enables

the model to perform superiorly on large-scale long-tailed and non-long-tailed datasets, which is a good starting point for mitigating the prevalent but unnoticed model bias.

****

DAG Matters! GFlowNets Enhanced Explainer for Graph Neural Networks

Wenqian Li,Yinchuan Li,Zhigang Li,Jianye HAO,Yan Pang

Uncovering rationales behind predictions of graph neural networks (GNNs) has received increasing attention over the years. Existing literature mainly focus on selecting a subgraph, through combinatorial optimization, to provide faithful explanations. However, the exponential size of candidate subgraphs limits the applicability of state-of-the-art methods to large-scale GNNs. We enhance on this through a different approach: by proposing a generative structure – GFlowNets-based GNN Explainer (GFlowExplainer), we turn the optimization problem into a step-by-step generative problem. Our GFlowExplainer aims to learn a policy that generates a distribution of subgraphs for which the probability of a subgraph is proportional to its’ reward. The proposed approach eliminates the influence of node sequence and thus does not need any pre-training strategies. We also propose a new cut vertex matrix to efficiently explore parent states for GFlowNets structure, thus making our approach applicable in a large-scale setting. We conduct extensive experiments on both synthetic and real datasets, and both qualitative and quantitative results show the superiority of our GFlowExplainer.

****

A MULTI-SCALE STRUCTURE-PRESERVING HETEROLOGOUS IMAGE TRANSFORMATION ALGORITHM BASED ON CONDITIONAL ADVERSARIAL NETWORK LEARNING

Rui Xiang,Guo-yo Wang,Pan Yu

Image transformation model learning is a basic technology for image enhancement, image super-resolution, image generation, multimodal image fusion, etc. which uses deep convolutional networks as a representation model for arbitrary functions, and uses fitting optimization with paired image training sets to solve the transformation model between images in the different sets. Affected by the complex and diverse changes of the 3D shape of the actual scene and the pixel-level optical properties of materials, the solution of the heterologous image conversion model is an ill-posed problem. In recent years, most of the proposed conditional adversarial learning methods for image transformation networks only consider the overall consistency loss constraint of the image, and the generated images often contain some pseudo-features or local structural deformations. In order to solve this problem, using the idea of multi-scale image coding and perception, this paper proposes a multi-scale structure-preserving heterologous image transformation method based on conditional adversarial network learning. First, using the idea of multi-scale coding and reconstruction, a multi-scale, step by step generator lightweight network structure is designed. Then, two image multi-scale structure loss functions are proposed, and combined with the existing overall consistency loss, a loss function for generative adversarial learning is designed. Finally, test experiments are performed on the KAIST-MPD-set1 dataset. The experimental results show that, compared with the state-of-the-art algorithms, the proposed algorithm can better suppress the local structural distortion, and has significant advantages in evaluation indicators such as RMSE, LPIPS, PSNR, and SSIM.

****

Metro: Memory-Enhanced Transformer for Retrosynthetic Planning via Reaction Tree

Songtao Liu,Zhitao Ying,Zuobai Zhang,Peilin Zhao,Jian Tang,Lu Lin,Dinghao Wu

Retrosynthetic planning plays a critical role in drug discovery and organic chemistry. Starting from a target molecule as the root node, it aims to find a complete reaction tree subject to the constraint that all leaf nodes belong to a set of starting materials. The multi-step reactions are crucial because they determine the flow chart in the production of the Organic Chemical Industry. However, existing datasets lack curation of tree-structured multi-step reactions and fail to provide such reaction trees, limiting models' understanding of organic molecule transformations. In this work, we first develop a benchmark curated for the retrosynthetic planning task, which consists of 124,869 reaction trees retrieved from the public USPTO-full dataset. On top of that, we propose Metro: Memory-Enh anced Transformer for RetrOsynthetic planning. Specifically, the dependency among molecules in the reaction tree is captured as context information for multi-step retrosynthesis predictions through transformers with a memory module. Extensive experiments show that Metro dramatically outperforms existing single-step retrosynthesis models by at least 10.7% in top-1 accuracy. The experiments demonstrate the superiority of exploiting context information in the retrosynthetic planning task. Moreover, the proposed model can be directly used for synthetic accessibility analysis, as it is trained on reaction trees with the shortest depths. Our work is the first step towards a brand new formulation for retrosynthetic planning in the aspects of data construction, model design, and evaluation.

****

In the ZONE: Measuring difficulty and progression in curriculum generation

Rose E Wang,Jesse Mu,Dilip Arumugam,Natasha Jaques,Noah Goodman

A common strategy in curriculum generation for reinforcement learning is to train a teacher network to generate tasks that fall within a student network's ``zone of proximal development'' (ZPD). These are tasks that are not too easy and not too hard for the student. Albeit intuitive, ZPD is not well understood computationally. We propose ZONE, a novel computational framework that operationalizes ZPD. It formalizes ZPD through the language of Bayesian probability theory, revealing that tasks should be selected by difficulty (the student's success probability on the task) and learning progression (the degree of change in the student's model parameters). ZONE operationalizes ZPD with two techniques that we apply on top of existing algorithms. One is REJECT, which rejects tasks outside a difficulty scope and the other is GRAD, which prioritizes tasks that maximize the student's gradient norm. Compared to the original algorithms, the ZONE techniques improve the student’s generalization performance on discrete Minigrid environments and continuous control Mujoco domains with up to $9 \times$ higher success. ZONE also accelerates the student's learning by training on up to $10\times$ less data.

****

Object Tracking by Hierarchical Part-Whole Attention

Jinkun Cao,Jiangmiao Pang,Xinshuo Weng,Rawal Khirodkar,Kris M. Kitani

We present in this paper that hierarchical representations of objects can provide an informative and low-noisy proxy to associate objects of interest in multi-object tracking. This is aligned with our intuition that we usually only need to compare a little region of the body of target objects to distinguish them from other objects. We build the hierarchical representation in levels of (1) target body parts, (2) the whole target body, and (3) the union area of the target and other objects of overlap. Furthermore, with the spatio-temporal attention mechanism by transformer, we can solve the tracking in a global fashion and keeps the process online. We design our method by combining the representation with the transformer and name it Hierarchical Part-Whole Attention, or HiPWA for short. The experiments on multiple datasets suggest its good effectiveness. Moreover, previous methods mostly focus on leveraging transformers to exploit long temporal context during association which requires heavy computation resources. But HiPWA focuses on a more informative representation of objects on every single frame instead. So it is more robust with the length of temporal context and more computationally economic.

****

Understanding the Training Dynamics in Federated Deep Learning via Aggregation Weight Optimization

Zexi Li,Tao Lin,Xinyi Shang,Chao Wu

From the server's perspective, federated learning (FL) learns a global model by iteratively sampling a cohort of clients and updating the global model with the sum local gradient of the cohort. We find this process is analogous to mini-batch SGD of centralized training. In mini-batch SGD, a model is learned by iteratively sampling a batch of data and updating the model with the sum gradient of the batch. In this paper, we delve into the training dynamics in FL by learning from the experience of optimization and generalization in mini-batch SGD. Specifically, we focus on two aspects: \emph{client coherence} (refers to sample coherenc e in mini-batch SGD) and \emph{global weight shrinking regularization} (refers to weight decay in mini-batch SGD). We find the roles of the two aspects are both determined by the aggregation weights assigned to each client during global model updating. Thus, we use aggregation weight optimization on the server as a tool to study how client heterogeneity and the number of local epochs affect the global training dynamics in FL. Besides, we propose an effective method for \textbf{Fed}erated \textbf{A}ggregation \textbf{W}eight \textbf{O}ptimization, named as \textsc{\textbf{FedAWO}}. Extensive experiments verify that our method can improve the generalization of the global model by a large margin on different datasets and models.

****

BiBench: Benchmarking and Analyzing Network Binarization

Haotong Qin,Mingyuan Zhang,Yifu Ding,Aoyu Li,Zhongang Cai,Ziwei Liu,Fisher Yu,Xianglong Liu

Neural network binarization emerges as one of the most promising compression approaches with extraordinary computation and memory savings by minimizing the bit-width of weight and activation. However, despite being a generic technique, recent works reveal that applying binarization in a wide range of realistic scenarios involving diverse tasks, architectures, and hardware is not trivial. Moreover, common challenges, such as severe degradation in accuracy and limited efficiency gains, suggest that specific attributes of binarization are not thoroughly studied and adequately understood. To close this gap, we present BiBench, a rigorously designed benchmark with in-depth analysis for network binarization. We first carefully scrutinize the requirements of binarization in the actual production setting. We thus define the evaluation tracks and metrics for a fair and systematic investigation. We then perform a comprehensive evaluation with a rich collection of milestone binarization algorithms. Our benchmark results show binarization still faces severe accuracy challenges but diminishing improvements brought by newer state-of-the-art binarization algorithms, even at the expense of efficiency. Moreover, the actual deployment of certain binarization operations reveals a surprisingly large deviation from their theoretical consumption. Finally, we provide suggestions based on our benchmark results and analysis, devoted to establishing a paradigm for accurate and efficient binarization among existing techniques. We hope BiBench paves the way towards more extensive adoption of network binarization and serves as a foundation for future research.

****

Contextual Image Masking Modeling via Synergized Contrasting without View Augmentation for Faster and Better Visual Pretraining

Shaofeng Zhang,Feng Zhu,Rui Zhao,Junchi Yan

We propose a new contextual masking image modeling (MIM) approach called contrasting-aided contextual MIM (ccMIM), under the MIM paradigm for visual pretraining. Specifically, we adopt importance sampling to select the masked patches with richer semantic information for reconstruction, instead of random sampling as done in previous MIM works. As such, the resulting patch reconstruction task from the remaining less semantic patches could be more difficult and helps to learn. To speed up the possibly slowed convergence due to our more difficult reconstruction task, we further propose a new contrastive loss that aligns the tokens of the vision transformer extracted from the selected masked patches and the remaining ones, respectively. The hope is that it serves as a regularizer for patch feature learning such that the image-level global information could be captured in both masked and unmasked patches, and notably such a single-view contrasting avoids the tedious image augmentation step required in recent efforts of introducing contrastive learning to MIM (to speedup convergence and discriminative ability). Meanwhile, the attention score from the contrastive global feature can also carry effective semantic clues to in turn guide our above masking patch selection scheme. In consequence, our contextual MIM and contrastive learning are synergetically performed in a loop (semantic patch selection-token alignment contrasting) to boost the best of the two worlds: fast convergence and strong performance on downstream tasks without ad-hoc augmentations, which are verified by empirical results on ImageNet-1K for both classification and dense vision tasks.

****

Patch-Level Contrasting without Patch Correspondence for Accurate and Dense Contrastive Representation Learning

Shaofeng Zhang,Feng Zhu,Rui Zhao,Junchi Yan

We propose ADCLR: \underline{A}ccurate and \underline{D}ense \underline{C}ontrastive \underline{R}epresentation \underline{L}earning, a novel self-supervised learning framework for learning accurate and dense vision representation. To extract spatial-sensitive information, ADCLR introduces query patches for contrasting in addition with global contrasting. Compared with previous dense contrasting methods, ADCLR mainly enjoys three merits: i) achieving both global-discriminative and spatial-sensitive representation, ii) model-efficient (no extra parameters in addition to the global contrasting baseline), and iii) correspondence-free and thus simpler to implement. Our approach achieves new state-of-the-art performance for contrastive methods. On classification tasks, for ViT-S, ADCLR achieves 78.1\% top-1 accuracy on ImageNet with linear probing, outperforming our baseline (DINO) without our devised techniques as plug-in, by 1.1\%. For ViT-B, ADCLR achieves 79.8\%, 84.0\% accuracy on ImageNet by linear probing and finetune, outperforming DINO by 0.6\%, 0.4\% accuracy. For dense tasks, on MS-COCO, ADCLR achieves significant improvements of 44.3\% AP on object detection, 39.7\% AP on instance segmentation, outperforming previous SOTA method SelfPatch by 2.2\% and 1.2\%, respectively. On ADE20K, ADCLR outperforms SelfPatch by 1.0\% mIoU, 1.2\% mAcc on the segmentation task.

****

Continuous-Discrete Convolution for Geometry-Sequence Modeling in Proteins

Hehe Fan,Zhangyang Wang,Yi Yang,Mohan Kankanhalli

The structure of proteins involves 3D geometry of amino acid coordinates and 1D sequence of peptide chains. The 3D structure exhibits irregularity because amino acids are distributed unevenly in Euclidean space and their coordinates are continuous variables. In contrast, the 1D structure is regular because amino acids are arranged uniformly in the chains and their sequential positions (orders) are discrete variables. Moreover, geometric coordinates and sequential orders are in two types of spaces and their units of length are incompatible. These inconsistencies make it challenging to capture the 3D and 1D structures while avoiding the impact of sequence and geometry modeling on each other. This paper proposes a Continuous-Discrete Convolution (CDConv) that uses irregular and regular approaches to model the geometry and sequence structures, respectively. Specifically, CDConv employs independent learnable weights for different regular sequential displacements but directly encodes geometric displacements due to their irregularity. In this way, CDConv significantly improves protein modeling by reducing the impact of geometric irregularity on sequence modeling. Extensive experiments on a range of tasks, including protein fold classification, enzyme reaction classification, gene ontology term prediction and enzyme commission number prediction, demonstrate the effectiveness of the proposed CDConv.

****

ELODI: Ensemble Logit Difference Inhibition for Positive-Congruent Training

Yue Zhao,Yantao Shen,Yuanjun Xiong,Shuo Yang,Wei Xia,Zhuowen Tu,Bernt Schiele,Stefano Soatto

Negative flips are errors introduced in a classification system when a legacy model is updated. Existing methods to reduce the negative flip rate (NFR) either do so at the expense of overall accuracy by forcing a new model to imitate the old models, or use ensembles, which multiply inference cost prohibitively. We analyze the role of ensembles in reducing NFR and observe that they remove negative flips that are typically not close to the decision boundary, but often exhibit large deviations in the distance among their logits. Based on the observation, we present a method, called Ensemble Logit Difference Inhibition ELODI, to train a classification system that achieves paragon performance in both error rate and NFR, at the inference cost of a single model. The method distills a homogeneous ensemble to a single student model which is used to update the classification system. ELODI also introduces a generalized distillation objective, Logit Difference Inhibition (LDI), which penalizes changes in the logits between the reference

ensemble and the student single model.
On multiple image classification benchmarks, model updates with ELODI demonstrate superior accuracy retention and NFR reduction.

****

Model-agnostic Measure of Generalization Difficulty

Akhilan Boopathy,Kevin Liu,Jaedong Hwang,Shu Ge,Asaad Mohammedsaleh,Ila R Fiete

The measure of a machine learning algorithm is the difficulty of the tasks it can perform, and sufficiently difficult tasks are critical drivers of strong machine learning models. However, quantifying the generalization difficulty of machine learning benchmarks has remained challenging. We propose what is to our knowledge the first model-agnostic measure of the inherent generalization difficulty of tasks. Our inductive bias complexity measure quantifies the total information required to generalize well on a task minus the information provided by the data. It does so by measuring the fractional volume occupied by hypotheses that generalize on a task given that they fit the training data. It scales exponentially with the intrinsic dimensionality of the space over which the model must generalize but only polynomially in resolution per dimension, showing that tasks which require generalizing over many dimensions are drastically more difficult than tasks involving more detail in fewer dimensions. Our measure can be applied to compute and compare supervised learning, reinforcement learning and meta-learning task difficulties against each other. We show that applied empirically, it formally quantifies intuitively expected trends, e.g. that in terms of required inductive bias, MNIST $<$ CIFAR10 $<$ Imagenet and fully observable Markov decision processes (MDPs) $<$ partially observable MDPs. Further, we show that classification of complex images $<$ few-shot meta-learning with simple images. Our measure provides a quantitative metric to guide the construction of more complex tasks requiring greater inductive bias, and thereby encourages the development of more sophisticated architectures and learning algorithms with more powerful generalization capabilities.

****

Efficient Exploration via Fragmentation and Recall

Jaedong Hwang,Zhang-Wei Hong,Eric R Chen,Akhilan Boopathy,Pulkit Agrawal,Ila R Fiete

Efficient exploration and model-building are critical for learning in large state- spaces. However, agents typically face problems like getting stuck locally during exploration and catastrophic forgetting in their construction of models when the environments are heterogeneous. Here, we propose and apply the concept of Fragmentation-and-Recall to solve spatial (FarMap) and reinforcement learning problems (FarCuriosity). Agents construct local maps or local models, respectively, which are used to predict the current observation. High surprisal points lead to a fragmentation event. At fracture points, we store the current map or model fragment in a long-term memory (LTM) and initialize a new fragment. On the other hand, Fragments are recalled (and thus reused) from LTM if the observations of their fracture points match the agent’s current observation during exploration. The set of fracture points defines a set of intrinsic potential subgoals. Agents choose their next subgoal from the set of near and far potential subgoals in the current fragment or LTM, respectively. Thus, local maps and model fragments guide exploration locally and avoid catastrophic forgetting in learning heterogeneous environments, while LTM promotes exploration more globally. We evaluate FarMap and FarCuriosity on complex procedurally-generated spatial environments and on reinforcement learning benchmarks and demonstrate that the proposed methods are more efficient at exploration and memory use, and in harvesting extrinsic rewards, respectively.

****

Hedge Your Actions: Flexible Reinforcement Learning for Complex Action Spaces

Norio Kosaka,Ayush Jain,Xinhu Li,Kyung-Min Kim,Joseph J Lim

Real-world decision-making is often associated with large and complex action representations, which can even be unsuited for the task. For instance, the items in recommender systems have generic representations that apply to each user differently, and the actuators of a household robot can be high-dimensional and noisy

. Prior works in discrete and continuous action space reinforcement learning (RL) define a retrieval-selection framework to deal with problems of scale. The retrieval agent outputs in the space of action representations to retrieve a few samples for a selection critic to evaluate. But, learning such retrieval actors becomes increasingly inefficient as the complexity in the action space rises. Thus, we propose to treat the retrieval task as one of listwise RL to propose a list of action samples that enable the selection phase to maximize the environment reward. By hedging its action proposals, we show that our agent is more flexible and sample efficient than conventional approaches while learning under a complex action space. Results are also present on \url{https://sites.google.com/view/complexaction}.

****