iclr 2023

Encoding Recurrence into Transformers

- Feiqing Huang, Kexin Lu, Yuxi CAI, Zhen Qin, Yanwen Fang, Guangjian Tian, Guodong Li
- abstract@open-review(Oral): 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.

Modeling content creator incentives on algorithm-curated platforms

- Jiri Hron, Karl Krauth, Michael Jordan, Niki Kilbertus, Sarah Dean
- abstract@open-review(Oral): 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.

Transfer NAS with Meta-learned Bayesian Surrogates

- Gresa Shala, Thomas Elsken, Frank Hutter, Josif Grabocka
- abstract@open-review(Oral): 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.

Scaling Up Probabilistic Circuits by Latent Variable Distillation

- Anji Liu, Honghua Zhang, Guy Van den Broeck
- abstract@open-review(Oral): 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.

A Kernel Perspective of Skip Connections in Convolutional Networks

- Daniel Barzilai, Amnon Geifman, Meirav Galun, Ronen Basri
- abstract@open-review(Oral): 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.

WikiWhy: Answering and Explaining Cause-and-Effect Questions

- Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, William Yang Wang
- abstract@open-review(Oral): 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-answer-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.

Git Re-Basin: Merging Models modulo Permutation Symmetries

- Samuel Ainsworth, Jonathan Hayase, Siddhartha Srinivasa
- abstract@open-review(Oral): 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 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 and CIFAR-100. 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 Role of Coverage in Online Reinforcement Learning

- Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, Sham M. Kakade
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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 classes. 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 a recently proposed classifier, the Vision Transformer, may have reached (or is about to reach) the Bayes error for the CIFAR-10H dataset.

Offline Q-learning on Diverse Multi-Task Data Both Scales And Generalizes

- Aviral Kumar, Rishabh Agarwal, Xinyang Geng, George Tucker, Sergey Levine
- abstract@open-review(Oral): 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.

What learning algorithm is in-context learning? Investigations with linear models

- Ekin Akyürek, Jacob Andreas, Dale Schuurmans, Tengyu Ma, Denny Zhou
- abstract@open-review(Oral): 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.

Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning

- Zeyuan Allen-Zhu, Yuanzhi Li
- abstract@open-review(Oral): (this is a theory paper)

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.

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
- abstract@open-review(Oral): 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.

Confidence-Conditioned Value Functions for Offline Reinforcement Learning

- Joey Hong, Aviral Kumar, Sergey Levine
- abstract@open-review(Oral): 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.

On the Sensitivity of Reward Inference to Misspecified Human Models

- Joey Hong, Kush Bhatia, Anca Dragan
- abstract@open-review(Oral): 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 both simulated biases, as well as real human data.

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
- abstract@open-review(Oral): 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 resolution 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.

Dichotomy of Control: Separating What You Can Control from What You Cannot

- Sherry Yang, Dale Schuurmans, Pieter Abbeel, Ofir Nachum
- abstract@open-review(Oral): 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 environment 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).

Learning where and when to reason in neuro-symbolic inference

- Cristina Cornelio, Jan Stuehmer, Shell Xu Hu, Timothy Hospedales
- abstract@open-review(Oral): 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 tasks and can further improve the performance of existing neuro-symbolic systems that lack our explicit reasoning during inference.

On the duality between contrastive and non-contrastive self-supervised learning

- Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, Yann LeCun
- abstract@open-review(Oral): 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 assumptions that contrastive and non-contrastive methods, respectively, need large batch sizes and 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.

DreamFusion: Text-to-3D using 2D Diffusion

- Ben Poole, Ajay Jain, Jonathan T. Barron, Ben Mildenhall
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching

- Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, Seunghoon Hong
- abstract@open-review(Oral): 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-shot 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.

Mitigating Gradient Bias in Multi-objective Learning: A Provably Convergent Approach

- Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, Tianyi Chen
- abstract@open-review(Oral): 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.

ReAct: Synergizing Reasoning and Acting in Language Models

- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, Yuan Cao
- abstract@open-review(Oral): 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 handle 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.

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
- abstract@open-review(Oral): 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 \oure, a conceptually and technically simple architecture that consists of three components: \circled{1} a \emph{link-encoder} that is only based on multi-layer perceptrons (MLP) to summarize the information from temporal links, \circled{2} a \emph{node-encoder} that is only based on neighbor mean-pooling to summarize node information, and \circled{3} an MLP-based \emph{link classifier} that performs link prediction based on the outputs of the encoders. Despite its simplicity, \our 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.

Is Conditional Generative Modeling all you need for Decision Making?

- Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, Pulkit Agrawal
- abstract@open-review(Oral): 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.

The Lie Derivative for Measuring Learned Equivariance

- Nate Gruver, Marc Anton Finzi, Micah Goldblum, Andrew Gordon Wilson
- abstract@open-review(Oral): 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 their

ability to directly encode translation equivariance 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 introduce 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.

Agree to Disagree: Diversity through Disagreement for Better Transferability

- Matteo Pagliardini, Martin Jaggi, François Fleuret, Sai Praneeth Karimireddy
- abstract@open-review(Oral): 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.

Efficient Conditionally Invariant Representation Learning

- Roman Pogodin, Namrata Deka, Yazhe Li, Danica J. Sutherland, Victor Veitch, Arthur Gretton
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

Transformers Learn Shortcuts to Automata

- Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, Cyril Zhang
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning

- Antonia Creswell, Murray Shanahan, Irina Higgins
- abstract@open-review(Oral): 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.

Compressing multidimensional weather and climate data into neural networks

- Langwen Huang, Torsten Hoefler
- abstract@open-review(Oral): 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 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.

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
- abstract@open-review(Oral): 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 fairness 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.

Near-optimal Coresets for Robust Clustering

- Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou, Xuan Wu
- abstract@open-review(Oral): 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.

Targeted Hyperparameter Optimization with Lexicographic Preferences Over Multiple Objectives

- Shaokun Zhang, Feiran Jia, Chi Wang, Qingyun Wu
- abstract@open-review(Oral): We propose to do targeted hyperparameter optimization with lexicographic preference over multiple objectives, motivated by various practical applications. We first provide a rigorous problem formulation. The formulation is novel and general to allow a clear specification of an automatable optimization goal. 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.

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
- abstract@open-review(Oral): 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.

Efficient Attention via Control Variates

- Lin Zheng, Jianbo Yuan, Chong Wang, Lingpeng Kong
- abstract@open-review(Oral): 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 approximation 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.

SAM as an Optimal Relaxation of Bayes

- Thomas Möllenhoff, Mohammad Emtiyaz Khan
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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 large language models and training GNNs on big graphs. 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 to separately train the two modules but at the same time allows 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.

Extreme Q-Learning: MaxEnt RL without Entropy

- Divyansh Garg, Joey Hejna, Matthieu Geist, Stefano Ermon
- abstract@open-review(Oral): 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) inspired by 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 (MaxEnt) 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. Finally, our method obtains strong results in the Offline D4RL benchmark outperforming prior works by 10-20 points on some tasks while offering moderate improvements over SAC and TD3 on online DM Control 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

- abstract@open-review(Oral): 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.

Simplified State Space Layers for Sequence Modeling

- Jimmy T.H. Smith, Andrew Warrington, Scott Linderman
- abstract@open-review(Oral): 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 achieving state-of-the-art performance on several long-range sequence modeling tasks. S5 averages 87.3% on the long range arena benchmark, and 98.5% on the most difficult Path-X task.

Moving Forward by Moving Backward: Embedding Action Impact over Action Semantics

- Kuo-Hao Zeng, Luca Weihs, Roozbeh Mottaghi, Ali Farhadi
- abstract@open-review(Oral): 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 environment and show that our AAP is highly performant even when faced, at inference-time, with missing actions and, previously unseen, perturbed action spaces. We will make the code and models for this work publicly available.

SimPer: Simple Self-Supervised Learning of Periodic Targets

- Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, Daniel McDuff
- abstract@open-review(Oral): 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.

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 Zhai, Neil Houlsby, Radu Soricut
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

Temporal Domain Generalization with Drift-Aware Dynamic Neural Networks

- Guangji Bai, Chen Ling, Liang Zhao
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

REVISITING PRUNING AT INITIALIZATION THROUGH THE LENS OF RAMANUJAN GRAPH

- Duc N.M Hoang, Shiwei Liu, Radu Marculescu, Zhangyang Wang
- abstract@open-review(Oral): 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 on different scores. 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 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 highly draconian assumption on the upper bound of the third largest eigenvalues (ˆµ) of layers belonging to highly sparse networks. We hence propose Iterative Mean Difference of Bound (IMDB) as a mean to relax the ˆµ upper bound. Likewise, we also show there exists a lower bound for ˆµ, which we call the NormAlized Random Coefficient (NaRC), that give 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 subnetworks preserving better 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. Codes will be made available upon acceptance.

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
- abstract@open-review(Oral): 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 domain, 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. Code and the dataset will be released.

A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification

- Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, Till J. Bungert
- abstract@open-review(Oral): 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-life 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/kjdhfg/fd-shifts

Fast and Precise: Adjusting Planning Horizon with Adaptive Subgoal Search

- Michał Zawalski, Michał Tyrolski, Konrad Czechowski, Damian Stachura, Piotr Piękos, Tomasz Odrzygóźdź, Yuhuai Wu, Łukasz Kuciński, Piotr Miłoś
- abstract@open-review(Oral): Complex reasoning problems contain states that vary in the computational cost required to determine a good action plan. Taking 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, allowing to focus on feasible further subgoals. In this way, AdaSubS benefits from the efficiency of planning with longer subgoals and the fine control with the shorter 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 inequality proving benchmark INT.

Towards Open Temporal Graph Neural Networks

- Kaituo Feng, Changsheng Li, Xiaolu Zhang, JUN ZHOU
- abstract@open-review(Oral): 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 disentangled 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.

Relative representations enable zero-shot latent space communication

- Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, Emanuele Rodolà
- abstract@open-review(Oral): 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, distinct latent spaces typically differ by an unknown quasi-isometric transformation: that is, in each space, the distances between the encodings do not change. In this work, we propose to adopt pairwise similarities as an alternative data representation, that can be used to enforce the desired invariance without any additional training. We show how neural architectures can leverage these relative representations to guarantee, in practice, latent isometry invariance, 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).

Language Modelling with Pixels

- Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, Desmond Elliott
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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.

Symbolic Physics Learner: Discovering governing equations via Monte Carlo tree search

- Fangzheng Sun, Yang Liu, Jian-Xun Wang, Hao Sun
- abstract@open-review(Oral): 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.

Clean-image Backdoor: Attacking Multi-label Models with Poisoned Labels Only

- Kangjie Chen, Xiaoxuan Lou, Guowen Xu, Jiwei Li, Tianwei Zhang
- abstract@open-review(Oral): 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. Experimental 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.

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
- abstract@open-review(Oral): 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.

Image to Sphere: Learning Equivariant Features for Efficient Pose Prediction

- David Klee, Ondrej Biza, Robert Platt, Robin Walters
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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 and purposefully. 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 with respect to the sequence length. Our experiments on five benchmark datasets show that compared with state-of-the-art methods, MICN yields 18.2% and 24.5% relative improvements for multivariate and univariate time series, respectively. Code will be released soon.

Personalized Federated Learning with Feature Alignment and Classifier Collaboration

- Jian Xu, Xinyi Tong, Shao-Lun Huang
- abstract@open-review(Oral): 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.

From Play to Policy: Conditional Behavior Generation from Uncurated Robot Data

- Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, Lerrel Pinto
- abstract@open-review(Oral): 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 suite 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: cbet-anon.github.io

Visual Classification via Description from Large Language Models

- Sachit Menon, Carl Vondrick
- abstract@open-review(Oral): 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.

The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation

- Zihui Xue, Zhengqi Gao, Sucheng Ren, Hang Zhao
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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 that maps β to the optimal parameters using hypernetworks. 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 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.

Near-optimal Policy Identification in Active Reinforcement Learning

- Xiang Li, Viraj Mehta, Johannes Kirschner, Ian Char, Willie Neiswanger, Jeff Schneider, Andreas Krause, Ilija Bogunovic
- abstract@open-review(Oral): 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.

Conditional Antibody Design as 3D Equivariant Graph Translation

- Xiangzhe Kong, Wenbing Huang, Yang Liu
- abstract@open-review(Oral): 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 affinity optimization. Specifically, the relative improvement to baselines is about 23\% in antigen-binding CDR design and 34\% for affinity optimization.

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
- abstract@open-review(Oral): 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.

Tailoring Language Generation Models under Total Variation Distance

- Haozhe Ji, Pei Ke, Zhipeng Hu, Rongsheng Zhang, Minlie Huang
- abstract@open-review(Oral): The standard paradigm of neural language generation adopts maximum likelihood estimation (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.

Transformers are Sample-Efficient World Models

- Vincent Micheli, Eloi Alonso, François Fleuret
- abstract@open-review(Oral): 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 codebase at this https URL. For the review process, we provide the code and visualizations in the supplementary materials.

Statistical Efficiency of Score Matching: The View from Isoperimetry

- Frederic Koehler, Alexander Heckett, Andrej Risteski
- abstract@open-review(Oral): 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 maximum 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.

View Synthesis with Sculpted Neural Points

- Yiming Zuo, Jia Deng
- abstract@open-review(Oral): 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, the scene geometry is parameterized using neural implicit representations (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 100x 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 in a user-friendly way.

AutoGT: Automated Graph Transformer Architecture Search

- Zizhao Zhang, Xin Wang, Chaoyu Guan, Ziwei Zhang, Haoyang Li, Wenwu Zhu
- abstract@open-review(Oral): 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.

Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting

- Yunhao Zhang, Junchi Yan
- abstract@open-review(Oral): Multivariate time series (MTS) forecasting applies to many practical scenarios. Recently many deep models have been proposed for 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 state-of-the-arts.

Betty: An Automatic Differentiation Library for Multilevel Optimization

- Sang Keun Choe, Willie Neiswanger, Pengtao Xie, Eric Xing
- abstract@open-review(Oral): 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 \url{https://github.com/leopard-ai/betty}.

Sparse Q-Learning: Offline Reinforcement Learning with Implicit Value Regularization

- Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Victor Wai Kin Chan, Xianyuan Zhan
- abstract@open-review(Oral): 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 actions will suffer from errors due to distributional shift. The recent proposed \textit{In-sample Learning} paradigm (e.g., 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 a practical algorithm, which uses the same value regularization as CQL, but in a complete in-sample manner. Compared with IQL, we find that our algorithm introduces sparsity in learning the value function, we thus dub our method Sparse Q-learning (SQL). We verify the effectiveness of SQL on D4RL benchmark datasets. We also show the benefits of sparsity by comparing SQL with IQL in noisy data

regimes and show the robustness of in-sample learning by comparing SQL with CQL in small data regimes. Under all settings, SQL achieves better results and owns faster convergence compared to other baselines.

Win: Weight-Decay-Integrated Nesterov Acceleration for Adaptive Gradient Algorithms

- Pan Zhou, Xingyu Xie, Shuicheng YAN
- abstract@open-review(Oral): Training deep networks on increasingly large-scale datasets is computationally challenging. In this work, we explore the problem of ``\textit{how to accelerate the convergence of adaptive gradient algorithms in a general manner}", and aim at providing practical insights to boost the training efficiency. To this end, we propose an effective and general {Weight-decay-Integrated Nesterov acceleration} (Win) for adaptive algorithms to enhance their convergence speed. 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 while fixing the above dynamic regularization brought by PPM. In this way, we arrive at our Win acceleration (like Nesterov 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 this 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 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 Transformer backbones. We hope Win acceleration shall be a default acceleration option for all popular optimizers in deep learning community to improve the training efficiency.

Towards Stable Test-time Adaptation in Dynamic Wild World

- Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, Mingkui Tan
- abstract@open-review(Oral): 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.

MocoSFL: enabling cross-client collaborative self-supervised learning

- Jingtao Li, Lingjuan Lyu, Daisuke Iso, Chaitali Chakrabarti, Michael Spranger
- abstract@open-review(Oral): 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. Thus, compared to the state-of-the-art FL-based approach, MocoSFL has significant advantages in both accuracy and practicality for cross-client applications.

Benchmarking Deformable Object Manipulation with Differentiable Physics

- Siwei Chen, Cunjun Yu, Yiqing Xu, Linfeng Li, Xiao Ma, Zhongwen Xu, David Hsu
- abstract@open-review(Oral): Deformable Object Manipulation (DOM) is of significant importance to both daily and industrial applications. Recent successes in differentiable physics simulators allow learning algorithms to train a policy with analytic gradients through environment dynamics, which significantly facilitates the development of DOM algorithms. However, existing DOM benchmarks are either single-object-based or non-differentiable. This leaves the questions of 1) how a task-specific algorithm performs on other tasks and 2) how a differentiable-physics-based algorithm compares with the non-differentiable ones in general. In this work, we present DaXBench, a differentiable DOM benchmark with a wide object and task coverage. DaXBench includes 9 challenging high-fidelity simulated tasks, covering rope, cloth, and liquid manipulation with various difficulty levels. To better understand the performance of general algorithms on different DOM tasks, we conduct comprehensive experiments over representative DOM methods, ranging from planning to imitation learning and reinforcement learning. In addition, we provide careful empirical studies of existing decision-making algorithms based on differentiable physics, and discuss their limitations, as well as potential future directions.

3D generation on ImageNet

- Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, Sergey Tulyakov
- abstract@open-review(Oral): All existing 3D-from-2D generators are designed for well-curated and alignable datasets: objects can be placed in the same position, similarly scaled and oriented, such that the camera always points to the center of the scene. This alignment procedure is infeasible for diverse, in-the-wild datasets: 1) it requires expensive annotation for each object category; and 2) most images are inherently "non-alignable" (e.g., it is impossible to align a "cat face" with a "kitchen"). As a result, existing 3D generators are not scalable to large in-the-wild datasets. In this work, for the first time, we propose a 3D generator which works on non-aligned datasets. First, we develop a technique to use an off-the-shelf, imprecise depth estimator to incorporate the 3D inductive bias into a GAN-based generator. Then, we create a novel learnable camera parametrization which does not use any alignment assumptions and construct a camera gradient penalty regularization. Finally, we propose a simple distillation-based technique to transfer the knowledge from an off-the-shelf feature embedder, like ResNet50, into our discriminator. Our work is the first one which develops a 3D generator for non-aligned data. We conduct experiments on SDIP Dogs, SDIP Elephants, LSUN Horse, and ImageNet on the 256x256 resolution to demonstrate the effectiveness of our ideas. Visualizations: https://u2wjb9xxz9q.github.io.

Rethinking the Expressive Power of GNNs via Graph Biconnectivity

- Bohang Zhang, Shengjie Luo, Liwei Wang, Di He
- abstract@open-review(Oral): 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.

Sparse Mixture-of-Experts are Domain Generalizable Learners

- Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, Ziwei Liu
- abstract@open-review(Oral): 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.

Token Merging: Your ViT But Faster

- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, Judy Hoffman
- abstract@open-review(Oral): 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.

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
- abstract@open-review(Oral): 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. Goal-directed 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 goal-directed 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.

Image as Set of Points

- Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, Yun Fu
- abstract@open-review(Oral): 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.

Generating Intuitive Fairness Specifications for Natural Language Processing

- Florian E. Dorner, Momchil Peychev, Nikola Konstantinov, Naman Goel, Elliott Ash, Martin Vechev
- abstract@open-review(Spotlight): 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.

On the Certification of Classifiers for Outperforming Human Annotators

- Qiongkai Xu, Christian Walder, Chenchen Xu
- abstract@open-review(Spotlight): 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 this 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.

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

- Jayaram Raghuram, Yijing Zeng, Dolores Garcia, Rafael Ruiz, Somesh Jha, Joerg Widmer, Suman Banerjee
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Learning Group Importance using the Differentiable Hypergeometric Distribution

- Thomas M. Sutter, Laura Manduchi, Alain Ryser, Julia E Vogt
- abstract@open-review(Spotlight): 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.

Concept-level Debugging of Part-Prototype Networks

- Andrea Bontempelli, Stefano Teso, Katya Tentori, Fausto Giunchiglia, Andrea Passerini
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 methods 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.

Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data

- Spencer Frei, Gal Vardi, Peter Bartlett, Nathan Srebro, Wei Hu
- abstract@open-review(Spotlight): 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.

Guarded Policy Optimization with Imperfect Online Demonstrations

- Zhenghai Xue, Zhenghao Peng, Quanyi Li, Zhihan Liu, Bolei Zhou
- abstract@open-review(Spotlight): Teacher-Student Framework (TSF) is a reinforcement learning setting where a teacher agent or human expert guards the training of a student agent by intervening and providing online demonstrations. Assuming the teacher policy is optimal, it has the perfect timing and capability to intervene the control 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 lower-bound safety guarantee without being affected by the teacher's own performance. Experiments on autonomous driving simulation show that our method can exploit teacher policies at any performance level and maintain a low training cost. Moreover, the student policy excels the imperfect teacher policy in terms of higher accumulated reward in held-out testing environments.

Learning with Logical Constraints but without Shortcut Satisfaction

- Zenan Li, Zehua Liu, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, Jian L\"{u}
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): We propose the novel certified training method, SABR, which outperforms existing methods across perturbation magnitudes on MNIST, CIFAR-10, and TinyImageNet, in terms of both standard and certifiable accuracies. The key insight behind SABR is 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. SABR does not only establish a new state-of-the-art in all commonly used benchmarks but more importantly, points to a new class of certified training methods promising to overcome the robustness-accuracy trade-off.

Multi-Objective Online Learning

- Jiyan Jiang, Wenpeng Zhang, Shiji Zhou, Lihong Gu, Xiaodong Zeng, Wenwu Zhu
- abstract@open-review(Spotlight): 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.

Seeing Differently, Acting Similarly: Heterogeneously Observable Imitation Learning

- Xin-Qiang Cai, Yao-Xiang Ding, Zixuan Chen, Yuan Jiang, Masashi Sugiyama, Zhi-Hua Zhou
- abstract@open-review(Spotlight): 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 Closer Look at Model Adaptation using Feature Distortion and Simplicity Bias

- Puja Trivedi, Danai Koutra, Jayaraman J. Thiagarajan
- abstract@open-review(Spotlight): Advances in the expressivity of large-scale 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, during FT have been found to achieve improved out-of-distribution generalization. A popular example is the recent LP+FT protocol which first learns a linear probe and then uses that initialization during FT. However, in this paper, we find that when adaptation protocols are also evaluated on a variety of safety objectives (e.g., calibration, robustness etc.), that a complementary perspective to feature distortion is required explain protocol behavior. To this end, we study the susceptibility of protocols to simplicity bias (SB), i.e. the well-known propensity of neural networks to rely upon simple features, as SB has recently been shown to underlie several problems in robust generalization. Using a synthetic dominoes dataset obtained by pairing (complex) CIFAR10 with (simple) MNIST samples, we demonstrate that the susceptibility of existing protocols to SB. Given the strong effectiveness of LP+FT, we propose incorporating hardness-promoting perturbations during LP to obtain initializations for FT that further decrease SB. We verify the effectiveness of these modified LP+FT protocols by decreasing SB on the dominoes dataset, and jointly improving OOD generalization and safety on standard adaptation benchmarks.

Understanding and Adopting Rational Behavior by Bellman Score Estimation

- Kuno Kim, Stefano Ermon
- abstract@open-review(Spotlight): 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.

STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

- Jaehyun Nam, Jihoon Tack, Kyungmin Lee, Hankook Lee, Jinwoo Shin
- abstract@open-review(Spotlight): 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 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 tasks. 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.

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
- abstract@open-review(Spotlight): 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 "perfect prompt" for a task. To mitigate the high degree of effort involved in prompting, we instead ask whether collecting multiple "imperfect prompts" and aggregating them can lead to a high quality prompting strategy. Our observations motivate our proposed prompting 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 ("Output True or False"). Our approach recursively uses the LLM itself to transform task inputs to the effective QA format. We apply these prompts to collect several noisy votes for the input's true label. We find that these prompts can have very different accuracies and complex dependencies and thus propose to use weak supervision, a classical procedure for combining the noisy predictions, to produce the final predictions. We evaluate AMA across open-source GPT-model families (e.g., Neo, BLOOM, OPT, and T0) demonstrating an average performance lift of 10.2% over the few-shot baseline across both small and large language models. This simple strategy enables the open-source GPT-Neo-6B model to match and exceed the performance of few-shot ($k \geq 32$) GPT3-175B on 15 of 20 popular benchmarks. Averaged across these tasks, the GPT-Neo-6B model outperforms few-shot GPT3-175B.

On Representing Linear Programs by Graph Neural Networks

- Ziang Chen, Jialin Liu, Xinshang Wang, Wotao Yin
- abstract@open-review(Spotlight): 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.

Scale-invariant Bayesian Neural Networks with Connectivity Tangent Kernel

- SungYub Kim, Sihwan Park, Kyung-Su Kim, Eunho Yang
- abstract@open-review(Spotlight): Explaining generalizations and preventing over-confident predictions are central goals of studies on the loss landscape of neural networks. Flatness, defined as loss invariability on perturbations of a pre-trained solution, is widely accepted as a predictor of generalization in this context. However, the problem that flatness and generalization bounds can be changed arbitrarily according to the scale of a parameter was pointed out, and previous studies partially solved the problem with restrictions: Counter-intuitively, their generalization bounds were still variant for the function-preserving parameter scaling transformation or limited only to an impractical network structure. As a more fundamental solution, we propose new prior and posterior distributions invariant to scaling transformations by \textit{decomposing} the scale and connectivity of parameters, thereby allowing the resulting generalization bound to describe the generalizability of a broad class of networks with the more practical class of transformations such as weight decay with batch normalization. We also show that the above issue adversely affects the uncertainty calibration of Laplace approximation and propose a solution using our invariant posterior. We empirically demonstrate our posterior provides effective flatness and calibration measures with low complexity in such a practical parameter transformation case, supporting its practical effectiveness in line with our rationale.

Minimalistic Unsupervised Learning with the Sparse Manifold Transform

- Yubei Chen, Zeyu Yun, Yi Ma, Bruno Olshausen, Yann LeCun
- abstract@open-review(Spotlight): We describe a minimalistic and interpretable method for unsupervised learning, without resorting to data augmentation, hyperparameter tuning, or other engineering designs, that achieves performance close to the 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 sparse manifold transform, one can 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 a simple gray-scale augmentation, the model gets 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 the SOTA methods. Additionally, we provide visualization to explain 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 there remains a small performance gap 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 learning.

GEASS: Neural causal feature selection for high-dimensional biological data

- Mingze Dong, Yuval Kluger
- abstract@open-review(Spotlight): 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 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 multi-dimensional 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.

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

- Sheng Li, Geng Yuan, Yue Dai, Youtao Zhang, Yanzhi Wang, Xulong Tang
- abstract@open-review(Spotlight): 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 manually pre-training defined freeze configurations, which do 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.

The In-Sample Softmax for Offline Reinforcement Learning

- Chenjun Xiao, Han Wang, Yangchen Pan, Adam White, Martha White
- abstract@open-review(Spotlight): 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 growing 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.

Unsupervised Meta-learning via Few-shot Pseudo-supervised Contrastive Learning

- Huiwon Jang, Hankook Lee, Jinwoo Shin
- abstract@open-review(Spotlight): 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.

Guiding Energy-based Models via Contrastive Latent Variables

- Hankook Lee, Jongheon Jeong, Sejun Park, Jinwoo Shin
- abstract@open-review(Spotlight): 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 propose 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.

Implicit regularization in Heavy-ball momentum accelerated stochastic gradient descent

- Avrajit Ghosh, He Lyu, Xitong Zhang, Rongrong Wang
- abstract@open-review(Spotlight): 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.

Real-time variational method for learning neural trajectory and its dynamics

- Matthew Dowling, Yuan Zhao, Il Memming Park
- abstract@open-review(Spotlight): 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.

Energy-Inspired Self-Supervised Pretraining for Vision Models

- Ze Wang, Jiang Wang, Zicheng Liu, Qiang Qiu
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Evolve Smoothly, Fit Consistently: Learning Smooth Latent Dynamics For Advection-Dominated Systems

- Zhong Yi Wan, Leonardo Zepeda-Nunez, Anudhyan Boral, Fei Sha
- abstract@open-review(Spotlight): 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.

BC-IRL: Learning Generalizable Reward Functions from Demonstrations

- Andrew Szot, Amy Zhang, Dhruv Batra, Zsolt Kira, Franziska Meier
- abstract@open-review(Spotlight): 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.

Dynamical systems embedding with a physics-informed convolutional network

- Matt Ricci, Noa Moriel, Zoe Piran, Mor Nitzan
- abstract@open-review(Spotlight): Dynamical systems are found in innumerable forms across the physical and biological sciences, yet all these systems fall naturally into universal equivalence classes: conservative or dissipative, stable or unstable, compressible or incompressible. Predicting these classes from data remains an essential open challenge in computational physics at which existing time-series classification methods struggle. Here, we propose, \texttt{phase2vec}, an embedding method that learns high-quality, physically-meaningful representations of dynamical systems without supervision. Our embeddings are produced by a convolutional backbone that extracts geometric features from flow data and is trained to minimize a physically-informed vector field reconstruction loss. The trained architecture can not only predict the equations of unseen data, but also, crucially, learns embeddings that respect the underlying semantics of the embedded physical systems. We first validate the quality of these learned embeddings by showing that the dynamical features they encode can be used to denoise corrupted testing data. Next, we examine the extent to which the underlying physical categories of input data can be decoded from embeddings compared to standard blackbox classifiers and state-of-the-art time series classification techniques. We find that our embeddings encode important physical properties of the underlying data, including the stability of fixed points, conservation of energy, and the incompressibility of flows, with greater fidelity than competing methods. We finally 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.

gDDIM: Generalized denoising diffusion implicit models

- Qinsheng Zhang, Molei Tao, Yongxin Chen
- abstract@open-review(Spotlight): 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 the 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.

FedExP: Speeding up Federated Averaging via Extrapolation

- Divyansh Jhunjhunwala, Shiqiang Wang, Gauri Joshi
- abstract@open-review(Spotlight): 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.

Serving Graph Compression for Graph Neural Networks

- Si Si, Felix Yu, Ankit Singh Rawat, Cho-Jui Hsieh, Sanjiv Kumar
- abstract@open-review(Spotlight): 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 whole 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 smaller number of nodes. In serving time, one just need 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.

Learning MLPs on Graphs: A Unified View of Effectiveness, Robustness, and Efficiency

- Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, Nitesh Chawla
- abstract@open-review(Spotlight): 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.

Contrastive Audio-Visual Masked Autoencoder

- Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, James R. Glass
- abstract@open-review(Spotlight): 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.

The Asymmetric Maximum Margin Bias of Quasi-Homogeneous Neural Networks

- Daniel Kunin, Atsushi Yamamura, Chao Ma, Surya Ganguli
- abstract@open-review(Spotlight): In this work, we explore the maximum-margin bias of quasi-homogeneous neural networks 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.

Optimal Transport for Offline Imitation Learning

- Yicheng Luo, zhengyao jiang, Samuel Cohen, Edward Grefenstette, Marc Peter Deisenroth
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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. 2020) 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 evaluation.

Learning multi-scale local conditional probability models of images

- Zahra Kadkhodaie, Florentin Guth, Stéphane Mallat, Eero P Simoncelli
- abstract@open-review(Spotlight): Deep neural networks can learn powerful prior probability models for images, as evidenced by high-quality synthesis results achieved with VAEs, GANs, and recent score-based diffusion methods. But these models are implicit, and 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 generalize a multi-scale model class motivated by the renormalization group of theoretical physics. It circumvents the curse of dimensionality by assuming Markov structure of multi-scale wavelet coefficients conditioned on coarser scale coefficients. We parameterize the model using conditional convolutional neural networks with local receptive fields, which enforce stationary Markov properties. We test the capabilities of the model on a dataset of face images, which are highly non-stationary and contain long-range geometric structures. Remarkably, denoising, super-resolution, and image synthesis results demonstrate that these structures are well-captured, using significantly smaller neighborhoods than required by a CNN operating in the pixel domain.

Disentangling with Biological Constraints: A Theory of Functional Cell Types

- James C. R. Whittington, Will Dorrell, Surya Ganguli, Timothy Behrens
- abstract@open-review(Spotlight): 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 disentangling 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, when, and how neurons represent factors in both brains and machines, and is a first step towards understanding of how task demands structure neural representations.

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

- abstract@open-review(Spotlight): 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.

Implicit Bias of Large Depth Networks: a Notion of Rank for Nonlinear Functions

- Arthur Jacot
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

Rhino: Deep Causal Temporal Relationship Learning with History-dependent Noise

- Wenbo Gong, Joel Jennings, Cheng Zhang, Nick Pawlowski
- abstract@open-review(Spotlight): 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.

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation

- Lorenz Kuhn, Yarin Gal, Sebastian Farquhar
- abstract@open-review(Spotlight): 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.

DINO as a von Mises-Fisher mixture model

- Hariprasath Govindarajan, Per Sidén, Jacob Roll, Fredrik Lindsten
- abstract@open-review(Spotlight): 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 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.

Associative Memory Augmented Asynchronous Spatiotemporal Representation Learning for Event-based Perception

- Uday Kamal, Saurabh Dash, Saibal Mukhopadhyay
- abstract@open-review(Spotlight): We propose $\textit{EventFormer}$, a computationally efficient event-based representation learning framework for asynchronously processing event camera data. EventFormer 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.

SMART: Self-supervised Multi-task pretrAining with contRol Transformers

- Yanchao Sun, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, Ashish Kapoor
- abstract@open-review(Spotlight): 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 work, 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.

TEMPERA: Test-Time Prompt Editing via Reinforcement Learning

- Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, Joseph E. Gonzalez
- abstract@open-review(Spotlight): 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.

Provable Defense Against Geometric Transformations

- Rem Yang, Jacob Laurel, Sasa Misailovic, Gagandeep Singh
- abstract@open-review(Spotlight): 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. Our results across multiple datasets 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.

Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations

- Polina Kirichenko, Pavel Izmailov, Andrew Gordon Wilson
- abstract@open-review(Spotlight): 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.

Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes

- Eoin M. Kenny, Mycal Tucker, Julie Shah
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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, surprisingly, 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.

Task-customized Masked Autoencoder via Mixture of Cluster-conditional Experts

- Zhili LIU, Kai Chen, Jianhua Han, Lanqing HONG, Hang Xu, Zhenguo Li, James Kwok
- abstract@open-review(Spotlight): 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, named Mixture of Cluster-conditional Experts (MoCE), which can be trained once but provide customized pre-training models for diverse downstream tasks. Different from the mixture of experts (MoE), our MoCE trains each expert only with semantically relevant images by using cluster-conditional gates, which can automatically format the hierarchical cluster architecture of the pre-training data. Thus, each downstream task can be allocated to its customized model pre-trained with data most similar to the downstream data. Experimental results show that our MoCE achieves 3.49\% average performance improvement compared to the vanilla MAE on a collection of 11 different downstream tasks.

Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization

- Jivat Neet Kaur, Emre Kiciman, Amit Sharma
- abstract@open-review(Spotlight): 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 results 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.

Using Language to Extend to Unseen Domains

- Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E. Gonzalez, Aditi Raghunathan, Anna Rohrbach
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

Hebbian Deep Learning Without Feedback

- Adrien Journé, Hector Garcia Rodriguez, Qinghai Guo, Timoleon Moraitis
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics

- Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, Sara Hooker
- abstract@open-review(Spotlight): Modern machine learning research relies on relatively few carefully curated datasets. 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 work 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.

Proposal-Contrastive Pretraining for Object Detection from Fewer Data

- Quentin Bouniot, Romaric Audigier, Angelique Loesch, Amaury Habrard
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Choreographer: Learning and Adapting Skills in Imagination

- Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Alexandre Lacoste, Sai Rajeswar
- abstract@open-review(Spotlight): 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://doubleblind-repos.github.io/

Learning About Progress From Experts

- Jake Bruce, Ankit Anand, Bogdan Mazoure, Rob Fergus
- abstract@open-review(Spotlight): 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 will open-source the curated expert training data at publication time.

Learning Fair Graph Representations via Automated Data Augmentations

- Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, Na Zou
- abstract@open-review(Spotlight): 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.

Emergence of Maps in the Memories of Blind Navigation Agents

- Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S. Morcos, Dhruv Batra
- abstract@open-review(Spotlight): 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.

Spectral Augmentation for Self-Supervised Learning on Graphs

- Lu Lin, Jinghui Chen, Hongning Wang
- abstract@open-review(Spotlight): 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.

Provably Efficient Neural Offline Reinforcement Learning via Perturbed Rewards

- Thanh Nguyen-Tang, Raman Arora
- abstract@open-review(Spotlight): We propose a novel offline reinforcement learning (RL) algorithm, namely PEturbed-Reward Value Iteration (PERVI) which amalgamates the randomized value function idea with the pessimism principle. 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, PERVI implicitly obtains pessimism by simply perturbing the offline data for multiple times with carefully-designed i.i.d Gaussian noises to learn an ensemble of estimated state-action values and acting greedily to the minimum of the ensemble. The estimated state-action values are obtained via fitting a parametric model (e.g. neural networks) to the perturbed datasets using gradient descent. As a result, PERVI 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 the potentially large log covering number in the learning bound. We prove that PERVI yields a provable uncertainty quantifier with overparameterized neural networks and achieves an $\tilde{\mathcal{O}}\left(\frac{ \kappa H^{5/2} \tilde{d} }{\sqrt{K}} \right)$ sub-optimality 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 PERVI with an empirical evaluation in a wide set of synthetic and real-world datasets. To the best of our knowledge, PERVI is the first offline RL algorithm that is both provably and computationally efficient in general Markov decision processes (MDPs) with neural network function approximation.

Self-supervised learning with rotation-invariant kernels

- Léon Zheng, Gilles Puy, Elisa Riccietti, Patrick Perez, Rémi Gribonval
- abstract@open-review(Spotlight): 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.

Neuromechanical Autoencoders: Learning to Couple Elastic and Neural Network Nonlinearity

- Deniz Oktay, Mehran Mirramezani, Eder Medina, Ryan P Adams
- abstract@open-review(Spotlight): 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.

Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training

- Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, Amy Zhang
- abstract@open-review(Spotlight): 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's frozen representation 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.

Sublinear Algorithms for Kernel Matrices via Kernel Density Estimation

- Ainesh Bakshi, Piotr Indyk, Praneeth Kacham, Sandeep Silwal, Samson Zhou
- abstract@open-review(Spotlight): 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{sublinear} 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 evaluation over baselines for LRA and a $\textbf{41x}$ reduction in the graph size for spectral sparsification.

A Higher Precision Algorithm for Computing the 1-Wasserstein Distance

- Pankaj K Agarwal, Sharath Raghvendra, Pouyan Shirzadian, Rachita Sowle
- abstract@open-review(Spotlight): We consider the problem of computing the 1-Wasserstein distance $\mathcal{W}(\mu,\nu)$ between two d-dimensional discrete distributions μ and ν that are within the unit hypercube. Let A (resp. B) be the support of μ (resp. ν). There are several algorithms that estimate $\mathcal{W}(\mu,\nu)$ within an additive factor 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 $\mathcal{W}(\mu,\nu)$ to an additive factor of $\min{\varepsilon, (d\log_{\sqrt{d}/\varepsilon} n)\mathcal{W}(\mu,\nu)}$. For the special case where every point in $A \cup B$ 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 to $\min{\varepsilon, (d\log\log n)\mathcal{W}(\mu,\nu)}$ in $O(T(n, \varepsilon/d)\log\log n)$ time.

Revisiting adapters with adversarial training

- Sylvestre-Alvise Rebuffi, Francesco Croce, Sven Gowal
- abstract@open-review(Spotlight): 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.

UNICORN: A Unified Backdoor Trigger Inversion Framework

- Zhenting Wang, Kai Mei, Juan Zhai, Shiqing Ma
- abstract@open-review(Spotlight): 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://anonymous.4open.science/r/UNICORN-FA0E.

ExpressivE: A Spatio-Functional Embedding For Knowledge Graph Completion

- Aleksandar Pavlovic, Emanuel Sallinger
- abstract@open-review(Spotlight): Knowledge graphs are inherently incomplete. Therefore substantial research has been directed towards knowledge graph completion (KGC), i.e., predicting missing triples from the information represented in the knowledge graph (KG). Embedding models have yielded promising results for KGC, yet any current KGC embedding model is incapable of: (1) fully capturing vital inference patterns (e.g., composition), (2) capturing prominent logical rules 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 embedding model 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 models and even significantly outperforms them on WN18RR.

Localized Randomized Smoothing for Collective Robustness Certification

- Jan Schuchardt, Tom Wollschläger, Aleksandar Bojchevski, Stephan Günnemann
- abstract@open-review(Spotlight): 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 and 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 guarantees.

Learning Probabilistic Topological Representations Using Discrete Morse Theory

- Xiaoling Hu, Dimitris Samaras, Chao Chen
- abstract@open-review(Spotlight): 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 the first deep learning based method to learn topological/structural representations. We use discrete Morse theory and persistent homology to construct an 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 rather 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.

Model-based Causal Bayesian Optimization

- Scott Sussex, Anastasia Makarova, Andreas Krause
- abstract@open-review(Spotlight): 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.

Training language models for deeper understanding improves brain alignment

- Khai Loong Aw, Mariya Toneva
- abstract@open-review(Spotlight): 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.

Dual Algorithmic Reasoning

- Danilo Numeroso, Davide Bacciu, Petar Veličković
- abstract@open-review(Spotlight): 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.

A Primal-Dual Framework for Transformers and Neural Networks

- Tan Minh Nguyen, Tam Minh Nguyen, Nhat Ho, Andrea L. Bertozzi, Richard Baraniuk, Stanley Osher
- abstract@open-review(Spotlight): 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 classification.

Fisher-Legendre (FishLeg) optimization of deep neural networks

- Jezabel R Garcia, Federica Freddi, Stathi Fotiadis, Maolin Li, Sattar Vakili, Alberto Bernacchia, Guillaume Hennequin
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 changes 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 3D rotations of human joints, body shape, and position information, while adapting to a given image and learning priors between joint rotation angles from large-scale data. 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 achieves 42.0 mm error on the PA-MPJPE metric of the challenging 3DPW, outperforming state-of-the-art counterparts by a large margin.

Efficient recurrent architectures through activity sparsity and sparse back-propagation through time

- Anand Subramoney, Khaleelulla Khan Nazeer, Mark Schöne, Christian Mayr, David Kappel
- abstract@open-review(Spotlight): Recurrent neural networks (RNNs) are well suited for solving sequence tasks in resource-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.

Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow

- Xingchao Liu, Chengyue Gong, qiang liu
- abstract@open-review(Spotlight): 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 generation and translation, our method yields nearly straight flows that give high quality results even with \emph{a single Euler discretization step}. Code will be made publicly available.

Inequality phenomenon in l_{∞}-adversarial training, and its unrealized threats

- Ranjie Duan, YueFeng Chen, Yao Zhu, Xiaojun Jia, Rong Zhang, Hui Xue'
- abstract@open-review(Spotlight): 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.

Learning Diffusion Bridges on Constrained Domains

- Xingchao Liu, Lemeng Wu, Mao Ye, qiang liu
- abstract@open-review(Spotlight): 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 Ω, 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.

Unsupervised Semantic Segmentation with Self-supervised Object-centric Representations

- Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, Thomas Brox
- abstract@open-review(Spotlight): 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 (47.3 mIoU; +10.1 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.

Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning

- Hao He, Kaiwen Zha, Dina Katabi
- abstract@open-review(Spotlight): 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, 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. Our code will be publicly available upon publication.

Decompositional Generation Process for Instance-Dependent Partial Label Learning

- Congyu Qiao, Ning Xu, Xin Geng
- abstract@open-review(Spotlight): 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 instance-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.

Building a Subspace of Policies for Scalable Continual Learning

- Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, Roberta Raileanu
- abstract@open-review(Spotlight): 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.

Not All Tasks Are Born Equal: Understanding Zero-Shot Generalization

- Jing Zhou, Zongyu Lin, Yanan Zheng, Jian Li, Zhilin Yang
- abstract@open-review(Spotlight): 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. We design a shuffle experiment to further show that training on these QA tasks leads to better cross-task generalization in multi-task learning under various training/test task splits. These novel findings combined deepen our understanding about zero-generalization---training on certain tasks such as QA encodes general knowledge transferable to a wide range of tasks, which explains the improved zero-shot performance in recent progress. 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.

Solving Constrained Variational Inequalities via a First-order Interior Point-based Method

- Tong Yang, Michael Jordan, Tatjana Chavdarova
- abstract@open-review(Spotlight): 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.

Symmetric Pruning in Quantum Neural Networks

- Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, Dacheng Tao
- abstract@open-review(Spotlight): 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.

Minimum Variance Unbiased N:M Sparsity for the Neural Gradients

- Brian Chmiel, Itay Hubara, Ron Banner, Daniel Soudry
- abstract@open-review(Spotlight): 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.

Knowledge-in-Context: Towards Knowledgeable Semi-Parametric Language Models

- Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, Jianshu Chen
- abstract@open-review(Spotlight): 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 changing 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 kinds of knowledge types: commonsense, entity, event, dictionary, script and causal knowledges. For each input instance, the KiC model adaptively selects a knowledge type and retrieves the knowledge pieces that are most helpful. 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. This key observation inspires us to develop a novel algorithm for learning KiC with an instance-adaptive knowledge selector. As a knowledge-rich semi-parametric language model, KiC only needs a relatively smaller parametric part to achieve superior zero-shot performance on unseen tasks. For instance, KiC-large with 770M parameters easily outperforms a 3B model on several benchmarks; that is, KiC exhibits emergent abilities at a much smaller model scale compared to the fully-parametric models.

Mosaic Representation Learning for Self-supervised Visual Pre-training

- Zhaoqing Wang, Ziyu Chen, Yaqian Li, Yandong Guo, Jun Yu, Mingming Gong, Tongliang Liu
- abstract@open-review(Spotlight): 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.

FluidLab: A Differentiable Environment for Benchmarking Complex Fluid Manipulation

- Zhou Xian, Bo Zhu, Zhenjia Xu, Hsiao-Yu Tung, Antonio Torralba, Katerina Fragkiadaki, Chuang Gan
- abstract@open-review(Spotlight): 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. FluidLab and FluidEngine will be publicly available.

Flow Matching for Generative Modeling

- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, Matthew Le
- abstract@open-review(Spotlight): 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 state-of-the-art performance in terms of both likelihood and sample quality, and allows fast and reliable sample generation using off-the-shelf numerical ODE solvers.

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
- abstract@open-review(Spotlight): 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, plasticine, sand, Newtonian and non-Newtonian fluids, and demonstrate significant performance gain on most tasks.

CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks

- Tuomas Oikarinen, Tsui-Wei Weng
- abstract@open-review(Spotlight): 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, which are required for existing tools to succeed. We show that CLIP-Dissect provides more accurate descriptions than existing methods for last layer neurons where the ground-truth is available as well 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 four minutes.

Data Continuity Matters: Improving Sequence Modeling with Lipschitz Regularizer

- Eric Qu, Xufang Luo, Dongsheng Li
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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: http://repo.codegen-iclr.org.

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
- abstract@open-review(Spotlight): 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.

Re-calibrating Feature Attributions for Model Interpretation

- Peiyu Yang, NAVEED AKHTAR, Zeyi Wen, Mubarak Shah, Ajmal Saeed Mian
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): Many Dec-POMDPs admit a qualitatively diverse set of `reasonable'' joint policies. Diversity literature is concerned with generating these joint 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 to diverse policy generation for turn-based Dec-POMDPs with public actions, which 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 butreasonable" play styles for the card game Hanabi, as indicated by their non-sabotaging behaviour and the graceful degradation of their performance with ad-hoc partners. We open-source our agents so that they may be used as starting points for a test bed for future research on (ad-hoc) coordination.

Augmented Lagrangian is Enough for Optimal Offline RL with General Function Approximation and Partial Coverage

- Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, Jiantao Jiao
- abstract@open-review(Spotlight): Offline reinforcement learning (RL), which refers to decision-making from a previously-collected dataset of interactions, 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 for successfully solving the sample-based approximation of the MIS problem is ensuring that certain state 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 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 state occupancy validity" than "promoters of conservatism."

DocPrompting: Generating Code by Retrieving the Docs

- Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, Graham Neubig
- abstract@open-review(Spotlight): 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 relevant 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.

A System for Morphology-Task Generalization via Unified Representation and Behavior Distillation

- Hiroki Furuta, Yusuke Iwasawa, Yutaka Matsuo, Shixiang Shane Gu
- abstract@open-review(Spotlight): 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.

Progress measures for grokking via mechanistic interpretability

- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, Jacob Steinhardt
- abstract@open-review(Spotlight): Neural networks often exhibit emergent behavior in which qualitatively new capabilities that arise from scaling up the number of parameters, training data, or even 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.

PiFold: Toward effective and efficient protein inverse folding

- Zhangyang Gao, Cheng Tan, Stan Z. Li
- abstract@open-review(Spotlight): 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.

Planning Goals for Exploration

- Edward S. Hu, Richard Chang, Oleh Rybkin, Dinesh Jayaraman
- abstract@open-review(Spotlight): 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. Project 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

- abstract@open-review(Spotlight): 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.

Score-based Generative 3D Mesh Modeling

- Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, Weiyang Liu
- abstract@open-review(Spotlight): 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. Existing 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 regular graph structure of meshes and use a simple yet very effective generative modeling method to generate 3D meshes. Specifically, we represent meshes in a deformable tetrahedral grid, and then train a diffusion model on this direct parameterization. We demonstrate the effectiveness of our model on multiple generative tasks.

Partially Observable RL with B-Stability: Unified Structural Condition and Sharp Sample-Efficient Algorithms

- Fan Chen, Yu Bai, Song Mei
- abstract@open-review(Spotlight): 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 weakly 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.

Domain Generalization via Heckman-type Selection Models

- Hyungu Kahng, Hyungrok Do, Judy Zhong
- abstract@open-review(Spotlight): 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.

A CMDP-within-online framework for Meta-Safe Reinforcement Learning

- Vanshaj Khattar, Yuhao Ding, Bilgehan Sel, Javad Lavaei, Ming Jin
- abstract@open-review(Spotlight): 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.

Effects of Graph Convolutions in Multi-layer Networks

- Aseem Baranwal, Kimon Fountoulakis, Aukosh Jagannath
- abstract@open-review(Spotlight): 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.

Post-hoc Concept Bottleneck Models

- Mert Yuksekgonul, Maggie Wang, James Zou
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Neural Networks Efficiently Learn Low-Dimensional Representations with SGD

- Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, Murat A Erdogdu
- abstract@open-review(Spotlight): 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.

Does Zero-Shot Reinforcement Learning Exist?

- Ahmed Touati, Jérémy Rapin, Yann Ollivier
- abstract@open-review(Spotlight): 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.

Hyperbolic Deep Reinforcement Learning

- Edoardo Cetin, Benjamin Paul Chamberlain, Michael M. Bronstein, Jonathan J Hunt
- abstract@open-review(Spotlight): 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-scale Physics

- Tailin Wu, Takashi Maruyama, Qingqing Zhao, Gordon Wetzstein, Jure Leskovec
- abstract@open-review(Spotlight): 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-scale 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-scale 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, which allows LAMP to adapt to varying relative importance of error vs. computation tradeoff at inference time. We test 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 with up to 60.5\% error reduction, and is able to adaptively trade-off computation to improve long-term prediction error.

Where to Begin? Exploring the Impact of Pre-Training and Initialization in Federated

- John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, Michael Rabbat
- abstract@open-review(Spotlight): 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 the fact that client devices have 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. We empirically study the impact of starting from a pre-trained model in federated learning using four standard federated learning benchmark datasets. 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 that future work proposing and evaluating federated optimization methods evaluate the performance when starting from random and pre-trained initializations. We also believe this study raises several questions for further work on understanding the role of heterogeneity in federated optimization.

Parametrizing Product Shape Manifolds by Composite Networks

- Josua Sassen, Klaus Hildebrandt, Benedikt Wirth, Martin Rumpf
- abstract@open-review(Spotlight): 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.

Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning?

- Rui Wen, Zhengyu Zhao, Zhuoran Liu, Michael Backes, Tianhao Wang, Yang Zhang
- abstract@open-review(Spotlight): 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{indiscriminative features} (INF). The existence of indiscriminative features makes the poisoned data become less useful for training a model, no matter if AT is applied or not. In contrast, existing methods are limited to using perturbations as \textit{shortcuts}, which just override the actual image content during model training. We demonstrate that for attacking a CIFAR-10 AT model under a reasonable setting with $\epsilon_\mathrm{adv}=\epsilon_\mathrm{poi}=8/255$, our INF 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 INF 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 confirm the effectiveness of INF in poisoning standard models.

Learning with Stochastic Orders

- Carles Domingo-Enrich, Yair Schiff, Youssef Mroueh
- abstract@open-review(Spotlight): 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 \emph{convex} or \emph{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 \emph{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.

MEDFAIR: BENCHMARKING FAIRNESS FOR MEDICAL IMAGEING

- Yongshuo Zong, Yongxin Yang, Timothy Hospedales
- abstract@open-review(Spotlight): A multitude of work has shown that machine learning-based medical diagnosis systems 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, nine 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.

Neural Design for Genetic Perturbation Experiments

- Aldo Pacchiano, Drausin Wulsin, Robert A Barton, Luis Voloch
- abstract@open-review(Spotlight): 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.

Efficient Discrete Multi Marginal Optimal Transport Regularization

- Ronak Mehta, Jeffery Kline, Vishnu Suresh Lokhande, Glenn Fung, Vikas Singh
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 finds 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.

Quantifying Memorization Across Neural Language Models

- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, Chiyuan Zhang
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 distribution 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.

Differentially Private L_2-Heavy Hitters in the Sliding Window Model

- Jeremiah Blocki, Seunghoon Lee, Tamalika Mukherjee, Samson Zhou
- abstract@open-review(Spotlight): 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.

NTFields: Neural Time Fields for Physics-Informed Robot Motion Planning

- Ruiqi Ni, Ahmed H Qureshi
- abstract@open-review(Spotlight): 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.

ZiCo: Zero-shot NAS via inverse Coefficient of Variation on Gradients

- Guihong Li, Yuedong Yang, Kartikeya Bhardwaj, Radu Marculescu
- abstract@open-review(Spotlight): Neural Architecture Search (NAS) is widely used to automatically design 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 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) for multiple datasets (CIFAR10/100, ImageNet16-120). 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 on ImageNet within 0.4 GPU days.

Pink Noise Is All You Need: Colored Noise Exploration in Deep Reinforcement Learning

- Onno Eberhard, Jakob Hollenstein, Cristina Pinneri, Georg Martius
- abstract@open-review(Spotlight): 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.

STaSy: Score-based Tabular data Synthesis

- Jayoung Kim, Chaejeong Lee, Noseong Park
- abstract@open-review(Spotlight): 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.

A Unified Algebraic Perspective on Lipschitz Neural Networks

- Alexandre Araujo, Aaron J Havens, Blaise Delattre, Alexandre Allauzen, Bin Hu
- abstract@open-review(Spotlight): 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 manner. 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 outperforms previous approaches on natural and certified accuracy.

The Influence of Learning Rule on Representation Dynamics in Wide Neural Networks

- Blake Bordelon, Cengiz Pehlevan
- abstract@open-review(Spotlight): 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.

Few-shot Cross-domain Image Generation via Inference-time Latent-code Learning

- Arnab Kumar Mondal, Piyush Tiwary, Parag Singla, Prathosh AP
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): Training deep reinforcement learning (DRL) models usually require 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.

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
- abstract@open-review(Spotlight): Sparse Neural Networks (SNNs) have received voluminous attention predominantly due to growing computational and memory footprints of consistently exploding parameter 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 12 datasets, that accounts for capturing a wide-range of domain-specific knowledge. Our systemic evaluation of representative SOTA sparse algorithms reveals an important obscured observation: all of the SOTA sparse algorithms bluntly fail to perform on SMC-Bench, sometimes at significantly trivial sparsity as low as 5%, which sought immediate attention of sparsity community to reconsider the highly proclaimed benefits of SNNs. By incorporating these well-thought and diverse tasks, SMC-Bench is designed to favor and encourage the development of highly generalizable sparse algorithms. We plan to open-source SMC-Bench evaluation suite to encourage sparsity researchers and assist them in building next-generation sparse algorithms with the potential to generalize on complex and practical tasks.

Sparse MoE with Random Routing as the New Dropout: Training Bigger and Self-Scalable Models

- Tianlong Chen, Zhenyu Zhang, AJAY KUMAR JAISWAL, Shiwei Liu, Zhangyang Wang
- abstract@open-review(Spotlight): Exploiting scale to revamp information absorption has recently become central to the success of deep learning and transformers have become $\textit{de facto}$ choice achieving numerous breakthrough performances on many real-world applications. Despite their enormous success, gigantic transformers suffer not only from exorbitant computational and memory footprints during training but also from severe collapse as evidenced by a high degree of parameter redundancy. Recently proposed Sparsely-activated Mixture-of-Experts (SMoEs) models have shown promise to mitigate the issue of training efficiency, yet they have some critical limitations. In particular, SMoEs models are prone to $\textit{redundant experts}$ due to representational collapse and \textit{poor scalability during 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, our work focuses on $\textit{exploring the overlooked scalability bottleneck of SMoEs}$, to effectively benefit scaling large-scale transformers. To this end, we propose a new plug-and-play training framework, $\textbf{SMoE-Dropout}$ to enable scaling transformers to better accuracy in the full capacity setting without collapse. Specifically, SMoE-Dropout consists of a $\textit{randomly initialized and fixed}$ router network to activate experts and gradually increase their number as training progresses over time. SMoE-Dropout naturally provides a $\textbf{``self-slimmable"}$ property offering consistent boosted performance for transformers with an increase in activated experts during inference and downstream fine-tuning, subjected to resource availability. Our extensive experiments across diverse transformer architectures on a variety of tasks validate superior performance and substantial computation savings, compared to densely trained baselines with equivalent parameter counts. More precisely, our trained BERT outperforms their 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 will be publicly released.

Adversarial Training of Self-supervised Monocular Depth Estimation against Physical-World Attacks

- Zhiyuan Cheng, James Chenhao Liang, Guanhong Tao, Dongfang Liu, Xiangyu Zhang
- abstract@open-review(Spotlight): 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 depth 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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion

- Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, Junchi Yan
- abstract@open-review(Spotlight): 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

- abstract@open-review(Spotlight): 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.

Gradient-based optimization is not necessary for generalization in neural networks

- Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, Tom Goldstein
- abstract@open-review(Spotlight): 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 that 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 one is found 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 tractable, we find that these non-gradient optimizers achieve test accuracy comparable to SGD.

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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$. 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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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.

Martingale Posterior Neural Processes

- Hyungi Lee, Eunggu Yun, Giung Nam, Edwin Fong, Juho Lee
- abstract@open-review(Spotlight): 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.

On the Usefulness of Embeddings, Clusters and Strings for Text Generation Evaluation

- Tiago Pimentel, Clara Isabel Meister, Ryan Cotterell
- abstract@open-review(Spotlight): 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.

DEP-RL: Embodied Exploration for Reinforcement Learning in Overactuated and Musculoskeletal Systems

- Pierre Schumacher, Daniel Haeufle, Dieter Büchler, Syn Schmitt, Georg Martius
- abstract@open-review(Spotlight): 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 been 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.

The Generalized Eigenvalue Problem as a Nash Equilibrium

- Ian Gemp, Charlie Chen, Brian McWilliams
- abstract@open-review(Spotlight): 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.

EA-HAS-Bench: Energy-aware Hyperparameter and Architecture Search Benchmark

- Shuguang Dou, XINYANG JIANG, Cai Rong Zhao, Dongsheng Li
- abstract@open-review(Spotlight): 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.

MARS: Meta-learning as Score Matching in the Function Space

- Krunoslav Lehman Pavasovic, Jonas Rothfuss, Andreas Krause
- abstract@open-review(Spotlight): 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 knowledge 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
- abstract@open-review(Spotlight): 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.

VA-DepthNet: A Variational Approach to Single Image Depth Prediction

- Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, Luc Van Gool
- abstract@open-review(Spotlight): We introduce VA-DepthNet, a simple, effective, and accurate deep neural network approach for a single image depth prediction (SIDP) problem. The proposed approach advocate 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 submission, our method---labeled as VA-DepthNet, when tested on the KITTI official depth-prediction evaluation set, indexed second on the leader board, and our accuracy is top performing among the published method.

Prompt-to-Prompt Image Editing with Cross-Attention Control

- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, Daniel Cohen-or
- abstract@open-review(Spotlight): 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.

DiffEdit: Diffusion-based semantic image editing with mask guidance

- Guillaume Couairon, Jakob Verbeek, Holger Schwenk, Matthieu Cord
- abstract@open-review(Spotlight): 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.

Corrupted Image Modeling for Self-Supervised Visual Pre-Training

- Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, Furu Wei
- abstract@open-review(Spotlight): 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 randomly 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
- abstract@open-review(Spotlight): 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.

Exploring Temporally Dynamic Data Augmentation for Video Recognition

- Taeoh Kim, Jinhyung Kim, Minho Shim, Sangdoo Yun, Myunggu Kang, Dongyoon Wee, Sangyoun Lee
- abstract@open-review(Spotlight): 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.

A General Framework for Sample-Efficient Function Approximation in Reinforcement Learning

- Zixiang Chen, Chris Junchi Li, Huizhuo Yuan, Quanquan Gu, Michael Jordan
- abstract@open-review(Spotlight): 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.

Adversarial Attacks on Adversarial Bandits

- Yuzhe Ma, Zhijin Zhou

- abstract@open-review(Spotlight): 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.

Ensuring DNN Solution Feasibility for Optimization Problems with Linear Constraints

- Tianyu Zhao, Xiang Pan, Minghua Chen, Steven Low
- abstract@open-review(Spotlight): 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. 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 Adversary-Sample Aware training algorithm to improve its optimality performance. We apply the framework to develop DeepOPF+ for solving the essential DC optimal power flow problem 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.

Simple Yet Effective Graph Contrastive Learning for Recommendation

- Xuheng Cai, Chao Huang, Lianghao Xia, Xubin Ren
- abstract@open-review(Spotlight): Graph neural network (GNN) is a powerful learning approach for graph-based recommender systems. Recently, GNN intergrated with contrastive learning has shown superior performance with data augmentation for recommendation, with the aim of 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 semantic 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 that negatively impact the generality and robustness of CL-based recommenders. Our model exclusively utilizes singular value decomposition for contrastive augmentation, which enables the unconstrained structure refinement with global collaborative relation modeling. Experiments conducted on several benchmark datasets demonstrate that our method significantly improves the performance over state-of-the-arts. Further analyses show the superiority of LightGCL's robustness against data sparsity and popularity bias. The source code of our model is available at https://anonymous.4open.science/r/LightGCL/.

MIMT: Masked Image Modeling Transformer for Video Compression

- Jinxi Xiang, Kuan Tian, Jun Zhang
- abstract@open-review(Spotlight): 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.

Hungry Hungry Hippos: Towards Language Modeling with State Space Models

- Tri Dao, Daniel Y Fu, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, Christopher Re
- abstract@open-review(Spotlight): 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 propose 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 H3-attention model that retains two attention layers surprisingly outperforms Transformers on OpenWebText by 1.0 PPL. When trained on the Pile at small/medium scale (125M and 355M parameters), hybrid H3-attention language models display 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. Next, to improve the efficiency of training SSMs on modern hardware, we propose FlashFFTConv. FlashFFTConv 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. FlashFFTConv yields 2$\times$ speedup on the long-range arena benchmark and allows hybrid language models to generate text 1.6$\times$ faster than Transformers.

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
- abstract@open-review(Spotlight): 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.

Relational Attention: Generalizing Transformers for Graph-Structured Tasks

- Cameron Diao, Ricky Loynd
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

Combinatorial-Probabilistic Trade-Off: P-Values of Community Properties Test in the Stochastic Block Models

- Shuting Shen, Junwei Lu
- abstract@open-review(Spotlight): 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.

Continuized Acceleration for Quasar Convex Functions in Non-Convex Optimization

- Jun-Kun Wang, Andre Wibisono
- abstract@open-review(Spotlight): 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.

Learning Soft Constraints From Constrained Expert Demonstrations

- Ashish Gaurav, Kasra Rezaee, Guiliang Liu, Pascal Poupart
- abstract@open-review(Spotlight): 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 and real world highway driving data.

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
- abstract@open-review(Spotlight): 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. Curiously, 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. Experiments across both language and vision transformers demonstrate that our LEarning to GrOw (LEGO) approach can save around 50% computational cost of training from scratch, while also consistently outperforming strong baselines that also reuse smaller pretrained models to initialize larger models. Our code will be made publicly available.

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
- abstract@open-review(Spotlight): 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.

UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks

- Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, Aniruddha Kembhavi
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

Learning to Estimate Shapley Values with Vision Transformers

- Ian Connick Covert, Chanwoo Kim, Su-In Lee
- abstract@open-review(Spotlight): 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 understanding 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 for generating 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.

A framework for benchmarking Class-out-of-distribution detection and its application to ImageNet

- Ido Galil, Mohammed Dabbah, Ran El-Yaniv
- abstract@open-review(Spotlight): 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 technique to benchmark image classifiers' ability to detect class-out-of-distribution instances (i.e., instances whose true labels the model does not recognize) at various levels of detection difficulty. We apply this technique to ImageNet, and benchmark 525 pretrained, publicly available, ImageNet-1k classifiers. We will provide the code to generate a benchmark for any ImageNet-1k classifier, along with the benchmarks prepared for the above-mentioned 525 models.

Additionally, we analyze the results from benchmarking these models and make numerous observations, including: (1) knowledge distillation consistently improves \emph{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.

Retrieval-based Controllable Molecule Generation

- Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, Anima Anandkumar
- abstract@open-review(Spotlight): 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.

Stochastic Multi-Person 3D Motion Forecasting

- Sirui Xu, Yu-Xiong Wang, Liangyan Gui
- abstract@open-review(Spotlight): 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 interaction, 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 movements 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 movement 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 show that our approach produces diverse and accurate multi-person predictions, significantly outperforming the state of the art.

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
- abstract@open-review(Spotlight): 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.

Sequential Latent Variable Models for Few-Shot High-Dimensional Time-Series Forecasting

- Xiajun Jiang, Ryan Missel, Zhiyuan Li, Linwei Wang

- abstract@open-review(Spotlight): 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 Charton, Gabriel Synnaeve
- abstract@open-review(Spotlight): 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.

Omnigrok: Grokking Beyond Algorithmic Data

- Ziming Liu, Eric J Michaud, Max Tegmark
- abstract@open-review(Spotlight): 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.

Flow Annealed Importance Sampling Bootstrap

- Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, José Miguel Hernández-Lobato
- abstract@open-review(Spotlight): 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 simulations, 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 where the flow is a poor approximation of the target, facilitating the discovery of new modes. We apply FAB to complex 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 samples with importance weights, we obtain unbiased histograms of dihedral angles that are almost identical to the ground truth.

Continual Unsupervised Disentangling of Self-Organizing Representations

- Zhiyuan Li, Xiajun Jiang, Ryan Missel, Prashnna Kumar Gyawali, Nilesh Kumar, Linwei Wang
- abstract@open-review(Spotlight): 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.

LMC: Fast Training of GNNs via Subgraph Sampling with Provable Convergence

- Zhihao Shi, Xize Liang, Jie Wang
- abstract@open-review(Spotlight): 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 subgraph-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.

Programmatically Grounded, Compositionally Generalizable Robotic Manipulation

- Renhao Wang, Jiayuan Mao, Joy Hsu, Hang Zhao, Jiajun Wu, Yang Gao
- abstract@open-review(Spotlight): 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 an input language instruction. Our framework uses a semantic parser to recover an executable program that is 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. The execution of the program 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}

SketchKnitter: Vectorized Sketch Generation with Diffusion Models

- Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, Yi-Zhe Song
- abstract@open-review(Spotlight): 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.

A Model or 603 Exemplars: Towards Memory-Efficient Class-Incremental Learning

- Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, De-Chuan Zhan
- abstract@open-review(Spotlight): 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.

Toeplitz Neural Network for Sequence Modeling

- Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, Yiran Zhong
- abstract@open-review(Spotlight): 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.

Learning QUBO Forms in Quantum Annealing

- Marcel Seelbach Benkner, Maximilian Krahn, Edith Tretschk, Zorah Lähner, Michael Moeller, Vladislav Golyanik
- abstract@open-review(Spotlight): Modern quantum annealers can find high-quality solutions to combinatorial optimization objectives given as quadratic unconstrained binary optimization (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 learned 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 will be open-sourced.

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
- abstract@open-review(Spotlight): 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.

On the complexity of nonsmooth automatic differentiation

- Jerome Bolte, Ryan Boustany, Edouard Pauwels, Béatrice Pesquet-Popescu
- abstract@open-review(Spotlight): 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.

Diffusion Posterior Sampling for General Noisy Inverse Problems

- Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, Jong Chul Ye

- abstract@open-review(Spotlight): 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.

Mass-Editing Memory in a Transformer

- Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, David Bau
- abstract@open-review(Spotlight): 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.

Learning the Positions in CountSketch

- Yi Li, Honghao Lin, Simin Liu, Ali Vakilian, David Woodruff
- abstract@open-review(Spotlight): 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.

Outcome-directed Reinforcement Learning by Uncertainty \& Temporal Distance-Aware Curriculum Goal Generation

- Daesol Cho, Seungjae Lee, H. Jin Kim
- abstract@open-review(Spotlight): 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.

A Laplace-inspired Distribution on SO(3) for Probabilistic Rotation Estimation

- Yingda Yin, Yang Wang, He Wang, Baoquan Chen
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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 attached in the supplementary materials and will be released to the public.

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
- abstract@open-review(Spotlight): Inspired by humans' remarkable ability to master arithmetic and generalize to unseen problems, we present a new dataset, HINT, to study machines' capability of learning generalizable concepts at three levels: perception, syntax, and semantics. In HINT, machines are tasked to learn 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. With a focus 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. We further design a few-shot learning split to test whether models could quickly learn new concepts and generalize them to more complex scenarios. To understand existing models' limitations, we conduct extensive experiments with various sequence-to-sequence models, including RNNs, Transformers, and GPT-3 (with the chain of thought prompting). The results suggest that current models still struggle in extrapolation to long-range syntactic dependency and semantics. Models show a significant gap toward human-level generalization when tested with new concepts in a few-shot setting. Moreover, we find that it is infeasible to solve HINT by simply scaling up the dataset and the model size; this strategy barely helps the extrapolation over syntax and semantics. Finally, in zero-shot GPT-3 experiments, the chain of thought prompting shows impressive results and significantly boosts the test accuracy. We believe the proposed dataset together with the experimental findings is of great interest to the community on systematic generalization.

Unsupervised Model Selection for Time Series Anomaly Detection

- Mononito Goswami, Cristian Ignacio Challu, Laurent Callot, Lenon Minorics, Andrey Kan

- abstract@open-review(Spotlight): 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 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, prediction error}, model centrality, and 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.

AANG : Automating Auxiliary Learning

- Lucio M. Dery, Paul Michel, Mikhail Khodak, Graham Neubig, Ameet Talwalkar
- abstract@open-review(Spotlight): Auxiliary objectives, supplementary learning signals that are introduced to help aid learning on data-starved or highly complex end-tasks, are commonplace in machine 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.

NeRN: Learning Neural Representations for Neural Networks

- Maor Ashkenazi, Zohar Rimon, Ron Vainshtein, Shir Levi, Elad Richardson, Pinchas Mintz, Eran Treister
- abstract@open-review(Spotlight): 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.

Formal Mathematics Statement Curriculum Learning

- Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, Ilya Sutskever
- abstract@open-review(Spotlight): 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.

Multifactor Sequential Disentanglement via Structured Koopman Autoencoders

- Nimrod Berman, Ilan Naiman, Omri Azencot
- abstract@open-review(Spotlight): 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 easy 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.

Packed Ensembles for efficient uncertainty estimation

- Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-marc Martinez, Andrei Bursuc, Gianni Franchi
- abstract@open-review(Spotlight): Deep Ensembles (DE) are a prominent approach 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 common backbone and forward pass to improve training and inference speeds. PE is designed to work under the memory budget of a single standard neural network. Through extensive studies we show that PE faithfully preserve the properties of DE, e.g., diversity, and match their performance in terms of accuracy, calibration, out-of-distribution detection and robustness to distribution shift.

Hidden Markov Transformer for Simultaneous Machine Translation

- Shaolei Zhang, Yang Feng
- abstract@open-review(Spotlight): 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. 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 we can only supervise the SiMT model with the observed target sequence. In this paper, we propose 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, used 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.

Multi-domain image generation and translation with identifiability guarantees

- Shaoan Xie, Lingjing Kong, Mingming Gong, Kun Zhang

- abstract@open-review(Spotlight): 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 provide the formulations of minimal changes and some other assumptions, under which the true joint distribution across domains is identifiable. We also provide a practical implementation of multi-domain image generation and a technique to improve unpaired image-to-image translation. 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.

Continual evaluation for lifelong learning: Identifying the stability gap

- Matthias De Lange, Gido M van de Ven, Tinne Tuytelaars
- abstract@open-review(Spotlight): 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 common state-of-the-art methods still suffer 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 of transient forgetting as the stability gap. The stability gap has 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 define a new set of metrics that enables quantifying the stability gap. Empirically we show that experience replay, constraint-based replay, and knowledge-distillation methods are all prone to the stability gap; and that the stability gap can be observed in both class 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 provide a conceptual explanation for the stability gap.

Domain-Indexing Variational Bayes for Domain Adaptation

- Zihao Xu, Hao He, Guang-Yuan Hao, Hao Wang
- abstract@open-review(Spotlight): 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.

One-Pixel Shortcut: On the Learning Preference of Deep Neural Networks

- Shutong Wu, Sizhe Chen, Cihang Xie, Xiaolin Huang
- abstract@open-review(Spotlight): 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, e.g., 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 in the same pixel that, if changed to a 0 or 1, could mostly and stably deviate from all original images. Since such calculation is only based on images, OPS needs significantly less computation than the previous methods based on model training. By 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.

Closing the gap: Exact maximum likelihood training of generative autoencoders using invertible layers

- Gianluigi Silvestri, Daan Roos, Luca Ambrogioni
- abstract@open-review(Spotlight): In this work, we provide an exact likelihood alternative to the variational training of generative autoencoders. This is achieved while leaving complete freedom in the choice of encoder, decoder and prior architectures, making our approach a drop-in replacement for the training of existing VAEs and VAE-style models. We refer to the resulting models as AutoEncoders within Flows (AEF), since the encoder, decoder and prior are defined as individual layers of an overall invertible architecture. This results in a deterministic encoding of the data, as opposed to the stochastic encoding of VAEs. We show that the approach often results in 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.

A Holistic View of Noise Transition Matrix in Deep Learning and Beyond

- LIN Yong, Renjie Pi, WEIZHONG ZHANG, Xiaobo Xia, Jiahui Gao, Xiao Zhou, Tongliang Liu, Bo Han
- abstract@open-review(Spotlight): 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 performance across multiple benchmarks.

Active Learning in Bayesian Neural Networks with Balanced Entropy Learning Principle

- Jae Oh Woo
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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 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.

GPViT: A High Resolution Non-Hierarchical Vision Transformer with Group Propagation

- Chenhongyi Yang, Jiarui Xu, Shalini De Mello, Elliot J. Crowley, Xiaolong Wang
- abstract@open-review(Spotlight): 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 together by a fixed number of learnable group tokens; we then perform Group Propagation where global information is exchanged between the grouped features; 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 segmentation, object detection, and instance segmentation. We show state-of-the-art performance across all tasks and significant gains over previous approaches on tasks requiring high- resolution outputs, for instance, our GPViT outperforms Swin Transformer-B by 2.0 mIoU on ADE20K semantic segmentation with only half as many parameters.

Neural Optimal Transport

- Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev
- abstract@open-review(Spotlight): 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 style translation task.

Dirichlet-based Uncertainty Calibration for Active Domain Adaptation

- Mixue Xie, Shuang Li, Rui Zhang, Chi Harold Liu
- abstract@open-review(Spotlight): 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. Extensive 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
- abstract@open-review(Spotlight): 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 an anonymous website \footnote{\href{https://anonymous.4open.science/r/ART_attention_retractable_transformer} {https://anonymous.4open.science/r/ART$_$attention$_$retractable$_$transformer}}.

Neural Episodic Control with State Abstraction

- Zhuo Li, Derui Zhu, Yujing Hu, Xiaofei Xie, Lei Ma, YAN ZHENG, Yan Song, Yingfeng Chen, Jianjun Zhao
- abstract@open-review(Spotlight): Existing Deep Reinforcement Learning (DRL) algorithms suffer from sample inefficiency. Generally, episodic control-based approaches are the solutions that leverage highly-rewarded past experiences to improve the DRL algorithms' sample efficiency. However, previous episodic control-based approaches fail to utilize the latent information of the historical behaviors (e,g., state transitions, topological similarities, etc.) and lack scalability during DRL training. This work introduces 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 environments in OpenAI gym domains. The experimental results indicate that NECSA can achieve better sample efficiency than the state-of-the-art episodic control-based approaches.

The Role of ImageNet Classes in Fréchet Inception Distance

- Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, Jaakko Lehtinen
- abstract@open-review(Spotlight): 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

- abstract@open-review(Spotlight): 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).

Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model

- Yinhuai Wang, Jiwen Yu, Jian Zhang
- abstract@open-review(Spotlight): 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.

Nonlinear Reconstruction for Operator Learning of PDEs with Discontinuities

- Samuel Lanthaler, Roberto Molinaro, Patrik Hadorn, Siddhartha Mishra
- abstract@open-review(Spotlight): 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 terms 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.

Learning Label Encodings for Deep Regression

- Deval Shah, Tor M. Aamodt
- abstract@open-review(Spotlight): Deep regression networks are widely used to tackle the problem of predicting a continuous value for a given input. Task-specialized approaches for training regression 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 \emph{Regularized Label Encoding Learning (RLEL)} for end-to-end training of an entire network and its label encodings. 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.

Multi-skill Mobile Manipulation for Object Rearrangement

- Jiayuan Gu, Devendra Singh Chaplot, Hao Su, Jitendra Malik
- abstract@open-review(Spotlight): 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.

Single-shot General Hyper-parameter Optimization for Federated Learning

- Yi Zhou, Parikshit Ram, Theodoros Salonidis, Nathalie Baracaldo, Horst Samulowitz, Heiko Ludwig
- abstract@open-review(Spotlight): 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.

Simplicial Embeddings in Self-Supervised Learning and Downstream Classification

- Samuel Lavoie, Christos Tsirigotis, Max Schwarzer, Ankit Vani, Michael Noukhovitch, Kenji Kawaguchi, Aaron Courville
- abstract@open-review(Spotlight): 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.

ViT-Adapter: Exploring Plain Vision Transformer for Accurate Dense Predictions

- Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, Yu Qiao
- abstract@open-review(Spotlight): 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. The code and models will be released.

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
- abstract@open-review(Spotlight): 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 accumulation 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.

Prompt Learning with Optimal Transport for Vision-Language Models

- Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, Kun Zhang
- abstract@open-review(Spotlight): 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.

DASHA: Distributed Nonconvex Optimization with Communication Compression and Optimal Oracle Complexity

- Alexander Tyurin, Peter Richtárik
- abstract@open-review(Spotlight): 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 ε-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.

LAVA: Data Valuation without Pre-Specified Learning Algorithms

- Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, Ruoxi Jia
- abstract@open-review(Spotlight): 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 $\textit{for free}$ 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 $\textit{orders of magnitude faster.}$

Meta-prediction Model for Distillation-Aware NAS on Unseen Datasets

- Hayeon Lee, Sohyun An, Minseon Kim, Sung Ju Hwang
- abstract@open-review(Spotlight): Distillation-aware Network Architecture Search (DaNAS) aims to search for an optimal student architecture that can obtain 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 a fixed source/target tasks and the teacher, which are not generalized well on a new task, thus need to perform 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 accuracies of the architectures obtained from the search. To eliminate excessive computational cost of DaNAS methods and the sub-optimality of rapid NAS methods, we propose a distillation-aware meta accuracy prediction model which can predict a given architecture's final performances on a dataset when performing KD with a given teacher, without having to actually train it on the target task. The experimental results demonstrate 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.

Denoising Diffusion Error Correction Codes

- Yoni Choukroun, Lior Wolf
- abstract@open-review(Spotlight): 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.

Exploring Active 3D Object Detection from a Generalization Perspective

- Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, Mahsa Baktashmotlagh
- abstract@open-review(Spotlight): 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 task, 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.

Neuro-Symbolic Procedural Planning with Commonsense Prompting

- Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, William Yang Wang
- abstract@open-review(Spotlight): Procedural planning aims to implement complex high-level goals by decomposition into sequential 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.

Generative Augmented Flow Networks

- Ling Pan, Dinghuai Zhang, Aaron Courville, Longbo Huang, Yoshua Bengio
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): Pre-trained representations (a.k.a. foundation models) have recently become a prevalent learning paradigm, where one first pre-trains a representation using large-scale unlabeled data, and then learns simple classifiers 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.

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
- abstract@open-review(Spotlight): 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 smooth, low-dimensional manifold 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.

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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): We present MBXP, an execution-based code completion benchmark in 10+ programming languages. This collection of datasets is generated by our conversion framework that translates prompts and test cases from the original MBPP dataset to the corresponding data in a target language. Based on this benchmark, we are able to evaluate code generation models in a multi-lingual fashion, and in particular discover generalization ability of language models on out-of-domain languages, advantages of large multi-lingual models over mono-lingual, benefits of few-shot prompting, 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. These solutions can be used for other code-related evaluations such as insertion-based, summarization, or code translation tasks where we demonstrate results and release as part of our benchmark.

GRACE-C: Generalized Rate Agnostic Causal Estimation via Constraints

- Mohammadsajad Abavisani, David Danks, Sergey Plis
- abstract@open-review(Spotlight): 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 algorithm 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.

Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs

- Yi-Lun Liao, Tess Smidt
- abstract@open-review(Spotlight): 3D-related inductive biases like translational invariance and rotational equivariance are indispensable to learning on 3D atomistic graphs such as molecules. Inspired by the success of Transformers in various domains, we study how to incorporate these inductive biases into Transformers. In this paper, we 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). Irreps features encode equivariant information in channels without complicating graph structures and thus enable us to directly incorporate them into Transformers by using equivariant operations. Moreover, we propose a novel attention mechanism called equivariant graph attention, which considers both content and geometric information contained in irreps features. The proposed equivariant graph attention improves upon typical attention in Transformers through replacing dot product attention with multi-layer perceptron attention and including non-linear message passing. We benchmark Equiformer on QM9, MD17, and OC20 datasets. Experiments demonstrate that Equiformer achieves competitive results to previous models and verify the effectiveness of the proposed attention.

MPCFORMER: FAST, PERFORMANT AND PRIVATE TRANSFORMER INFERENCE WITH MPC

- Dacheng Li, Hongyi Wang, Rulin Shao, Han Guo, Eric Xing, Hao Zhang
- abstract@open-review(Spotlight): Enabling private inference is crucial for many cloud inference services that are based on Transformer models. However, existing private inference solutions for Transformers can increase the inference latency by more than 60$\times$ or significantly compromise the quality of inference results. In this paper, we design the framework MPCFORMER using secure multi-party computation (MPC) and Knowledge Distillation (KD). It can be used in tandem with many specifically designed MPC-friendly approximations and trained Transformer models. We use MPCFORMER to significantly speed up Transformer model inference in MPC settings while achieving similar ML performance to the original model. We evaluate MPCFORMER with various settings in MPC. On the IMDb dataset, we achieve similar performance to $\text{BERT}\textit{BASE}$, *while being 5.3$\times$ faster. On the GLUE benchmark, we achieve 97% performance of $\textit{BERT}*\text{BASE}$ with a 2.2$\times$ speedup. We show that MPCFORMER remains effective with different trained Transformer weights such as $\text{ROBERTA}\textit{BASE}$ *and larger models including $\textit{BERT}*\text{LARGE}$. In particular, we achieve similar performance to BERT_LARGE, while being 5.93$\times$ faster on the IMDb dataset.

Disparate Impact in Differential Privacy from Gradient Misalignment

- Maria S. Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, Jesse C Cresswell
- abstract@open-review(Spotlight): 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 gradient clipping as the most significant source. This observation leads us to a new method for reducing unfairness by preventing gradient misalignment in DPSGD.

TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second

- Noah Hollmann, Samuel Müller, Katharina Eggensperger, Frank Hutter
- abstract@open-review(Spotlight): 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 30 datasets from the OpenML-CC18 suite, we show that our method clearly outperforms boosted trees and performs on par with complex state-of-the-art AutoML systems with up to $70\times$ speedup. This increases to a $3\,200\times$ speedup when a GPU is available. We provide all our code, the trained TabPFN, an interactive browser demo and a Colab notebook at https://github.com/tabpfn-anonym/TabPFNAnonym.

Human Motion Diffusion Model

- Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Amit Haim Bermano, Daniel Cohen-or
- abstract@open-review(Spotlight): 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.

Visual Recognition with Deep Nearest Centroids

- Wenguan Wang, Cheng Han, Tianfei Zhou, Dongfang Liu
- abstract@open-review(Spotlight): 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 will be released.

Continuous PDE Dynamics Forecasting with Implicit Neural Representations

- Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, patrick gallinari
- abstract@open-review(Spotlight): 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.

No Reason for No Supervision: Improved Generalization in Supervised Models

- Mert Bülent Sarıyıldız, Yannis Kalantidis, Karteek Alahari, Diane Larlus
- abstract@open-review(Spotlight): 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 while 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 analyse 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. Finally, we perform several analyses of the features and class weights to present insights on how each component of our setup affects the training and learned representations.

EVA3D: Compositional 3D Human Generation from 2D Image Collections

- Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, Ziwei Liu
- abstract@open-review(Spotlight): 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.

Voxurf: Voxel-based Efficient and Accurate Neural Surface Reconstruction

- Tong Wu, Jiaqi Wang, Xingang Pan, Xudong XU, Christian Theobalt, Ziwei Liu, Dahua Lin
- abstract@open-review(Spotlight): 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 will be made publicly available.

Generating Diverse Cooperative Agents by Learning Incompatible Policies

- Rujikorn Charakorn, Poramate Manoonpong, Nat Dilokthanakul
- abstract@open-review(Spotlight): 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

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
- abstract@open-review(Spotlight): 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.

ISS: Image as Stepping Stone for Text-Guided 3D Shape Generation

- Zhengzhe Liu, Peng Dai, Ruihui Li, XIAOJUAN QI, Chi-Wing Fu
- abstract@open-review(Spotlight): Text-guided 3D shape generation remains challenging due to the absence of large paired text-shape data, 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 text-guided shape stylization module to dress up the output shapes with novel 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 textur

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
- abstract@open-review(Spotlight): 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.

Neural Collapse Inspired Feature-Classifier Alignment for Few-Shot Class-Incremental Learning

- Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, Dacheng Tao
- abstract@open-review(Spotlight): 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. Our code will be publicly available.

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
- abstract@open-review(Spotlight): 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 (10250 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
- abstract@open-review(Spotlight): 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
- abstract@open-review(Spotlight): 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.

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
- abstract@open-review(Spotlight): 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. Our code will be released.

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
- abstract@open-review(Spotlight): 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 will be released for facilitating further research and application.

Minimax Optimal Kernel Operator Learning via Multilevel Training

- Jikai Jin, Yiping Lu, Jose Blanchet, Lexing Ying
- abstract@open-review(Spotlight): 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.

Sparse and Hierarchical Masked Modeling for Convolutional Representation Learning

- Keyu Tian, Yi Jiang, qishuai diao, Chen Lin, Liwei Wang, Zehuan Yuan
- abstract@open-review(Spotlight): This paper presents a simple yet powerful framework to pre-train convolutional network (convnet) with Sparse masKed modeling. SparK addresses key challenges in applying transformer-specialized masked modeling to convolutional models: (i) convolution operation cannot handle irregular, random-masked input; (ii) the single-scale nature of existing masked modeling is inconsistent with convnet's hierarchical structure. For (i), we sparsely gather the unmasked pixels to a sparse image and use sparse convolution for encoding. For the later, we develop a hierarchical encoder-decoder to reconstruct from multi-scale encoded features to fully exploit the advantage of hierarchy. As the first hierarchical masked modeling method designed for convnets, SparK exploits their untapped potential. On three downstream tasks, SparK surpasses both state-of-the-art contrastive learning and \textit{transformer-based} masked modeling by similarly large margins (around +1.0%). Improvements on object detection and instance segmentation are more substantial (>1.0%), verifying strong transferability of features learned by SparK. We also demonstrate SparK's favorable scaling behavior by observing more gains on larger models. Taken all results together, a promising future of generative pre-training on convnets has been initially shown by SparK. Codes will be made publicly available.

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

- Julius Adebayo, Melissa Hall, Bowen Yu, Bobbie Chern
- abstract@open-review(Poster): 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.

Factorized Fourier Neural Operators

- Alasdair Tran, Alexander Mathews, Lexing Xie, Cheng Soon Ong
- abstract@open-review(Poster): 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 several insights that collectively have a significant effect – the separable spectral representations; improved residual connections; and carefully designed training strategies. 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 Kolmogorov flow, 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
- abstract@open-review(Poster): Most structured pruning algorithms achieve subnetworks which not only have high predictive accuracy but also have significantly lower FLOPs. It is now noted that the decrease in FLOPs seldom results in a similar decrease in inference time. These algorithms avoid pruning coupled channels (CCs). These channels contribute significantly to the total inference time; layers with CCs as input or output take more than 66% of the inference time in ResNet-50. Motivated by this, we study the problem of pruning CCs in the data-free regime in this paper. Formal studies for pruning CCs are sparse due to a lack of proper characterization. Thus, we define Data Flow Couplings (DFCs) that abstract the notion of coupling and aid us in scoring coupled elements of the network. Gauging

saliencies of CCs is not straightforward, for there exists a discrepancy among the layerwise importance of CCs using conventional scoring strategies. This necessitates the definition of grouped saliencies to gauge the importance of coupled elements in a network. Since we do not have access to data, we propose the Backwards Graph-based Saliency Computation (BGSC) algorithm that computes saliencies by estimating an upper bound to the reconstruction error of intermediate layers. We then compare saliencies to prune CCs and call this pruning strategy DFPC. Finally, we show the efficacy of DFPC for models trained on CIFAR-10, CIFAR-100, and ImageNet datasets. For instance, we find that for a 5% accuracy drop and 1.64x reduction of FLOPs for ResNet-101 trained on the CIFAR-10 dataset, the inference time speedup obtained by DFPC is up to 1.66x, without finetuning. When assuming access to the ImageNet training set, we significantly improve over the data-free method. We see at least a 47.1% improvement 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
- abstract@open-review(Poster): 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 with only access to random samples drawn from the original distribution and without access to the original training set or loss function. We posit the following hypothesis: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%.

Adversarial Training descends without descent: Finding actual descent directions based on Danskin's theorem

- Fabian Latorre, Igor Krawczuk, Leello Tadesse Dadi, Thomas Pethick, Volkan Cevher
- abstract@open-review(Poster): 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 (DDD) 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.

Learning Continuous Normalizing Flows For Faster Convergence To Target Distribution via Ascent Regularizations

- Shuangshuang Chen, Sihao Ding, Yiannis Karayiannidis, Mårten Björkman
- abstract@open-review(Poster): 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. Although 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 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 cost. Furthermore, the flows show to stabilize themselves to mitigate performance deterioration and are less sensitive to the choice of training flow length T .

Softened Symbol Grounding for Neuro-symbolic Systems

- Zenan Li, Yuan Yao, Taolue Chen, Jingwei Xu, Chun Cao, Xiaoxing Ma, Jian L\"{u}
- abstract@open-review(Poster): Neuro-symbolic learning usually consists of two worlds, i.e., neural network learning and symbolic constraint satisfaction, whose effectiveness hinges on symbol grounding, a fundamental problem in AI. This paper presents a novel, softened symbol grounding process, enabling the interactions of the two worlds in a mutually beneficial manner. Technically, we design a neuro-symbolic learning framework that features (1) modeling of deterministic symbol solution states as a Boltzmann distribution, which avoids expensive state searching and facilitates the interaction between network training and symbolic reasoning; (2) an efficient MCMC sampling technique leveraging projection and SMT solvers, which overcomes the connectivity barrier in sampling symbol solution spaces; (3) an annealing mechanism that avoids the trap of sub-optimal symbol groundings. Experiments with three representative neuro-symbolic learning tasks demonstrate that, thanks to its superior symbol grounding capability, our framework successfully solves problems well beyond the frontier of the existing proposals.

Mini-batch k-means terminates within $O(d/\epsilon)$ iterations

- Gregory Schwartzman
- abstract@open-review(Poster): 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.

Finally, we show the applicability of our results to the mini-batch k-means algorithm implemented in the scikit-learn (sklearn) python library.

Learning Uncertainty for Unknown Domains with Zero-Target-Assumption

- Yu Yu, Hassan Sajjad, Jia Xu
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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).

QAID: Question Answering Inspired Few-shot Intent Detection

- Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, Boaz Carmeli
- abstract@open-review(Poster): Intent detection with semantically similar fine-grained intents is a challenging task. To address it, we reformulate intent detection as a question-answering 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 first stage, we train the model to learn better query representation in a self supervise manner. Then, in the second stage, we fine-tune the model to optimize 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.

Solving stochastic weak Minty variational inequalities without increasing batch size

- Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, Volkan Cevher
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): Co-design of morphology and control of a Voxel-based Soft Robot (VSR) for solving a given task is challenging due to the bi-level optimization in the enormous combined design and policy space. 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 stage of the curriculum, we use reinforcement learning to simultaneously train the design and policy, which is enabled by incorporating the design process into the environment and using differentiable policy representations. The converged morphology, the learned design and control policies from the last stage are inherited and serve as the starting point for the next stage. In the 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.

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

- Tribhuvanesh Orekondy, Pratik Kumar, Shreya Kadambi, Hao Ye, Joseph Soriaga, Arash Behboodi
- abstract@open-review(Poster): 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 oflocal' 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).

LS-IQ: Implicit Reward Regularization for Inverse Reinforcement Learning

- Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, Jan Peters
- abstract@open-review(Poster): 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 for improving stability, mistreating or even neglecting 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 the 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, 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.

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

- Zebang Shen, Jiayuan Ye, Anmin Kang, Hamed Hassani, Reza Shokri
- abstract@open-review(Poster): 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, CENTAUR 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.

EquiMod: An Equivariance Module to Improve Self-Supervised Learning

- Alexandre DEVILLERS, Mathieu Lefort

- abstract@open-review(Poster): Self-supervised visual representation methods are closing the gap with supervised learning performance. These 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. color is important for birds and flower classification). Few recent works proposed to mitigate the 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 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 SimCLR and BYOL, increases the performances on 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.

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

- Prashant Shivaram Bhat, Bahram Zonooz, Elahe Arani
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 significantly improve quality by 0.5-2%, possibly with some length increase.

FairGBM: Gradient Boosting with Fairness Constraints

- André Cruz, Catarina G Belém, João Bravo, Pedro Saleiro, Pedro Bizarro
- abstract@open-review(Poster): 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 performance guarantees and low cost. However, in these domains bias is a concern. Existing in-processing Fair ML methods are either inapplicable to GBDT, or incur in significant train time overhead, or are inadequate for problems with high class imbalance -- a typical issue in high stakes domains. 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 have to employ a ``proxy-Lagrangian'' formulation using smooth convex error rate proxies to enable gradient-based optimization. Our implementation shows an order of magnitude speedup in training time when compared with related work, a pivotal aspect to foster the widespread adoption of FairGBM by real-world practitioners.

Online Bias Correction for Task-Free Continual Learning

- Aristotelis Chrysakis, Marie-Francine Moens
- abstract@open-review(Poster): Task-free continual learning is the machine-learning setting where a model is trained online with data generated by a non-stationary 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 show both theoretically and empirically how experience replay biases the outputs of the model towards recent stream observations. Moreover, we propose a simple approach to correct for this bias online, by changing the way the output layer of the model is optimized. We show that our approach improves significantly the learning performance of experience-replay approaches over a number of different datasets. Our findings suggest that, in contrast to stationary machine-learning problems, the output layer of a model should be optimized separately from its preceding layers when performing experience replay.

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

- Hugo Schmutz, Olivier HUMBERT, Pierre-Alexandre Mattei
- abstract@open-review(Poster): 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.

Making Substitute Models More Bayesian Can Enhance Transferability of Adversarial Examples

- Qizhang Li, Yiwen Guo, Wangmeng Zuo, Hao Chen
- abstract@open-review(Poster): 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 16.8% absolute increase in average attack success rate on ImageNet), and, by combining with these recent methods, further performance gain can be obtained. Our code will be publicly available.

Cross-Layer Retrospective Retrieving via Layer Attention

- Yanwen Fang, Yuxi CAI, Jintai Chen, Jingyu Zhao, Guangjian Tian, Guodong Li
- abstract@open-review(Poster): 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.

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

- Shmuel Bar David, Itamar Zimerman, Eliya Nachmani, Lior Wolf
- abstract@open-review(Poster): 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

Unveiling the sampling density in non-uniform geometric graphs

- Raffaele Paolino, Aleksandar Bojchevski, Stephan Günnemann, Gitta Kutyniok, Ron Levie
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 these 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://anonymous.4open.science/r/ReScore-7631.

Iterative Circuit Repair Against Formal Specifications

- Matthias Cosler, Frederik Schmitt, Christopher Hahn, Bernd Finkbeiner
- abstract@open-review(Poster): 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.

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

- Mingxu Tao, Yansong Feng, Dongyan Zhao
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Offline reinforcement learning (RL) is a challenging setting where existing off-policy actor-critic methods perform poorly due to the overestimation of out-of-distribution actions. Thus, various additional augmentations are proposed to keep the learned policy close to the offline dataset (or behavior policy). In this work, starting from the analysis of offline monotonic policy improvement, we get a surprising finding that some 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 accomplish the closeness. Based on this, we design an algorithm called Behavior Proximal Policy Optimization (BPPO), which successfully solves offline RL without any extra constraint or regularization introduced. Extensive experiments on the D4RL benchmark indicate this extremely succinct method outperforms state-of-the-art offline RL algorithms.

Actionable Neural Representations: Grid Cells from Minimal Constraints

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

- abstract@open-review(Poster): 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 understanding 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.

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
- abstract@open-review(Poster): 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.

Geometrically regularized autoencoders for non-Euclidean data

- Cheongjae Jang, Yonghyeon Lee, Yung-Kyun Noh, Frank C. Park
- abstract@open-review(Poster): 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.

A Message Passing Perspective on Learning Dynamics of Contrastive Learning

- Yifei Wang, Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, Yisen Wang
- abstract@open-review(Poster): 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 function 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.

Zeroth-Order Optimization with Trajectory-Informed Derivative Estimation

- Yao Shu, Zhongxiang Dai, Weicong Sng, Arun Verma, Patrick Jaillet, Bryan Kian Hsiang Low
- abstract@open-review(Poster): 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 additional function queries are required for derivative estimation in their every GD update, which typically hinders their deployment in applications where every function query is expensive. To this end, we propose a trajectory-informed derivative estimation method which only uses the optimization trajectory (i.e., the history of function queries during optimization) and hence eliminates 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.

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

- Taiji Suzuki, Atsushi Nitanda, Denny Wu
- abstract@open-review(Poster): 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.

Asynchronous Distributed Bilevel Optimization

- Yang Jiao, Kai Yang, Tiancheng Wu, Dongjin Song, Chengtao Jian
- abstract@open-review(Poster): 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.

Confidence-Based Feature Imputation for Graphs with Partially Known Features

- Daeho Um, Jiwoong Park, Seulki Park, Jin young Choi
- abstract@open-review(Poster): 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.

LiftedCL: Lifting Contrastive Learning for Human-Centric Perception

- Ziwei Chen, Qiang Li, Xiaofeng Wang, Wankou Yang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 R\'enyi 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 R\'enyi 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
- abstract@open-review(Poster): Quality Diversity (QD) has emerged as a powerful alternative optimization paradigm that aims at generating and maintaining 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, for instance 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). Both lines of work demonstrated great promise but approaches based on BBO tend to be less sample-efficient as they often empirically evaluate gradients through random sampling in the parameter space. 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 RL algorithms that ME was so far 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 within a population update 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 also open source an efficient JAX-based implementation of our algorithm.

Gray-Box Gaussian Processes for Automated Reinforcement Learning

- Gresa Shala, André Biedenkapp, Frank Hutter, Josif Grabocka
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning in temporally structured environments

- Matt Jones, Tyler R. Scott, Mengye Ren, Gamaleldin Fathy Elsayed, Katherine Hermann, David Mayo, Michael Curtis Mozer
- abstract@open-review(Poster): Natural environments have temporal structure at multiple timescales, a property that is reflected in biological learning and memory but typically not in machine learning systems. This paper advances a multiscale learning model in which each weight in a neural network is decomposed into a sum

of subweights learning independently with different learning and decay rates. Thus knowledge becomes distributed across different timescales, enabling rapid adaptation to task changes while avoiding catastrophic interference with older knowledge. First, we prove that previous models that learn at multiple timescales, but with complex coupling between timescales, are formally equivalent to the multiscale learner via a reparameterization that eliminates this coupling. Thus the multiscale learning offers a unifying framework that is conceptually and computationally simpler than past work. The same analysis also offers a new characterization of momentum learning, as a fast weight with a negative learning rate. Second, we derive a model of Bayesian inference in environments governed by $1/f$ noise, a common pattern in both natural and human-generated environments that involves long-range (power law) autocorrelations. The model works by applying a Kalman filter to jointly infer dynamics at multiple timescales. We then derive a variational approximation to the Bayesian model and show that it is equivalent to the multiscale learner. Third, we evaluate the models in synthetic online prediction tasks characterized by $1/f$ noise in the latent parameters of the environment. We find that the Bayesian model significantly outperforms stochastic gradient descent (which effectively learns at only one timescale) and a batch heuristic that predicts each timestep based on a fixed horizon of past observations (motivated by the idea that older data have gone stale). Moreover, the multiscale learner with parameters obtained from 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 (vs. quadratic for the Bayesian model). Future work will incorporate the multiscale learner as an optimizer in deep networks to explore their ability to learn in rich temporally structured environments.

RandProx: Primal-Dual Optimization Algorithms with Randomized Proximal Updates

- Laurent Condat, Peter Richtárik
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Fast Nonlinear Vector Quantile Regression

- Aviv A. Rosenberg, Sanketh Vedula, Yaniv Romano, Alexander Bronstein
- abstract@open-review(Poster): $$ \newcommand{\rvar}[1]{\mathrm {#1}} \newcommand{\rvec}[1]{\boldsymbol{\mathrm{#1}}} $$ Quantile regression (QR) is a powerful tool for estimating one or more conditional 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.

Leveraging Large Language Models for Multiple Choice Question Answering

- Joshua Robinson, David Wingate
- abstract@open-review(Poster): 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 lengths 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 tasks and largely closes the gap with the SOTA, suggesting that the MCQA ability of LLMs has been previously underestimated.

Regression with Label Differential Privacy

- Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash Varadarajan, Chiyuan Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Object rearrangement is a challenge for embodied agents because solving these tasks requires generalizing across a combinatorially large set of underlying entities that take the value of object states. Worse, these entities are often 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 inputs. By constructing a factorized transition graph over clusters of object 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 a set of simulated rearrangement and stacking tasks.

Selective Frequency Network for Image Restoration

- Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, Alois Knoll
- abstract@open-review(Poster): Image restoration aims to reconstruct the latent sharp image from the corrupted observation. Besides dealing with this long-standing task in the spatial domain, a few approaches seek solutions in the frequency domain based on the large discrepancy between spectra of sharp/degraded image pairs. However, these works utilize the existing transformation tools, e.g., Wavelet Transform, to split the feature into several 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 the feature into separate frequency subbands dynamically and locally, and then accentuate the useful ones via the channel-wise attention mechanism. In addition, aiming to cope with the large-scale degradation kernel, we propose an extremely simple decoupling and modulation module to enlarge the receptive field based on 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 image dehazing, image motion/defocus deblurring, image desnowing, and image deraining.

Improving Differentiable Neural Architecture Search by Encouraging Transferability

- Parth Sheth, Pengtao Xie
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, GPUs, and customized neural network accelerators. Our experimental results on CPUs show that MA-BERT achieves up to 27% reduction in inference time with comparable accuracy on many downstream tasks compared to the baseline BERT models.

Efficient Certified Training and Robustness Verification of Neural ODEs

- Mustafa Zeqiri, Mark Niklas Mueller, Marc Fischer, Martin Vechev
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

CASR: Generating Complex Sequences with Autoregressive Self-Boost Refinement

- Hongwei Han, Mengyu Zhou, Shi Han, Xiu Li, Dongmei Zhang
- abstract@open-review(Poster): 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 classic 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%).

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
- abstract@open-review(Poster): Although training machine learning models for robustness is critical for real-world adoption, determining how to best ensure robustness remains an open problem. Some methods (e.g., DRO) are overly conservative, while others (e.g., Group DRO) require domain knowledge that may be hard to obtain. 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 is simple. For example, we may expect that group shifts occur along high-level features (e.g., image background, lighting). Thus, we aim to learn a model that maintains high accuracy on simple group functions realized by these features, but need not spend valuable model capacity achieving high accuracy on contrived groups of examples. Based on this idea, we formulate a two-player game where conditioned on the label the adversary can only separate datapoints into potential groups using simple features, which corresponds to a bitrate constraint on the adversary's capacity. Our resulting practical algorithm, Bitrate-Constrained DRO (BR-DRO), does not require group annotations on training data yet matches the performance of Group DRO on datasets that have them or are long-tailed. Our theoretical analysis reveals that in some settings BR-DRO objective can provably yield statistically efficient and less pessimistic solutions than unconstrained DRO.

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

- Matus Telgarsky
- abstract@open-review(Poster): 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 can not 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
- abstract@open-review(Poster): 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 margin 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, Chenzhong Yin, Gengshuo Liu, Radu Balan, Paul Bogdan
- abstract@open-review(Poster): 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 L2 error compared to the best results from the state-of-the-art models.

Mid-Vision Feedback for Convolutional Neural Networks

- Michael Maynord, Eadom T Dessalene, Cornelia Fermuller, Yiannis Aloimonos
- abstract@open-review(Poster): Feedback plays a prominent role in biological vision, where perception is modulated based on agents' continuous interactions with the world, and evolving expectations and world model. We introduce a novel mechanism which modulates perception in Convolutional Neural Networks (CNNs) 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 CNN. 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 - applications range from image and video understanding to explainable AI and robotics. 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.

Safe Reinforcement Learning From Pixels Using a Stochastic Latent Representation

- Yannick Hogewind, Thiago D. Simão, Tal Kachman, Nils Jansen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Intelligent neuron models in Natural Language Processing (NLP) are known to be vulnerable to textual Trojan attacks, i.e., Trojan models behave normally for normal inputs, yet produce malicious output for input with a trigger. Invisible textual triggers, e.g., syntactic-structure triggers, are becoming popular since they require more effort for detection and defense than triggers based on content insertion and replacement. Although the high stealthy and attack effects, current Trojan attacks with syntactic-structure triggers are highly dependent on a large corpus of training data to generate poisoned samples with the specific syntactic structure for Trojan insertion. Accessing training data for attackers is not always realistic, training-time attacks and current syntactic poisoned trigger generation, and Trojan insertion by updating all the parameters are extremely time-consuming. In this paper, we propose TrojText to study whether the

invisible textual Trojan attack can be efficiently performed without the presence of training data in a more realistic and cost-efficient manner. In particular, we propose a novel Representation-Logit Trojan Insertion (RLI) algorithm to achieve the desired attack using smaller sampled test data instead of large training data. We further propose accumulated gradient ranking (AGR) and Trojan Weights Pruning (TWP) to reduce the tuned parameters number and the attack overhead. We perform extensive experiments of AG's News, SST-2, and OLID on BERT and XLNet. Our TrojText could classify 98.35 \% of test sentences into target class on the BERT model for AG's News data.

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
- abstract@open-review(Poster): 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 PINNs is extremely sensitive to the interplay between different loss terms and hyper-parameters, including the learning rate. In this work, we propose a Bayesian approach through a 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 more than ten times faster than the PDE's solution. The evaluation outcome shows that our proposed method is more robust for solving the EIT problem.

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

- Yunchong Song, Chenghu Zhou, Xinbing Wang, Zhouhan Lin
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 eight competing SOTA defense methods show the empirical superiority of our method on both single-shot and continuous FL backdoor attacks. We will release our code upon publication.

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
- abstract@open-review(Poster): 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 languages 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 an improved and refreshed variant of the mC4 multilingual corpus consisting of 29 trillion characters across 107 languages. In addition we release full code to reproduce our experiments.

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

- Xiaoqi Wang, Han Wei Shen
- abstract@open-review(Poster): 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 with the existing work, GNNInterpreter is more computationally efficient and more flexible in generating explanation graphs with different types of node features and edge features, without introducing another blackbox to explain the GNN and without requiring manually specified domain-specific knowledge. Additionally, the experimental studies conducted on four different datasets demonstrate that the explanation graph generated by GNNInterpreter can match the desired graph pattern when the model is ideal and reveal potential model pitfalls if there exist any.

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

- Kei Sen Fong, Shelvia Wongso, Mehul Motani
- abstract@open-review(Poster): 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 2 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
- abstract@open-review(Poster): 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, few studies comprehensively explore the representation capability of distinct pre-trained language models on different antibody problems. Here, to investigate the problem, we aim to answer the following key questions: (1) How do pre-trained language models perform in antibody tasks with different specificity? (2) How many benefits will the model gain if we introduce the specific biological mechanism to the pretraining process? (3) Do the learned antibody pre-trained representations make sense in real-world antibody problems, like drug discovery and immune process understanding? Previously, no benchmark available largely hindered the study to answer these questions. To facilitate the investigation, we provide an AnTibody Understanding Evaluation (ATUE) benchmark. We comprehensively evaluate the performance of protein pretrained language models by empirical study along with conclusions and new insights.

Learning to reason over visual objects

- Shanka Subhra Mondal, Taylor Whittington Webb, Jonathan Cohen
- abstract@open-review(Poster): A core component of human intelligence is the ability to identify abstract patterns governing 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. This 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 enable abstract visual reasoning. We found that a simple model, consisting only of an object-centric encoder and a transformer reasoning module, displayed performance approaching state-of-the-art methods on two challenging RPM-like benchmarks (PGM and I-RAVEN), suggesting that an inductive bias for object-centric processing is a key component of visual reasoning, and may supplant some problem-specific inductive biases.

Imitating Graph-Based Planning with Goal-Conditioned Policies

- Junsu Kim, Younggyo Seo, Sungsoo Ahn, Kyunghwan Son, Jinwoo Shin
- abstract@open-review(Poster): 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.

A theoretical study of inductive biases in contrastive learning

- Jeff Z. HaoChen, Tengyu Ma
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Computational Language Acquisition with Theory of Mind

- Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, Graham Neubig
- abstract@open-review(Poster): 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.1 We model ToM by giving the speaker agent an internal listener model that is trained alongside the speaker and using this ToM model to rerank potential utterances. We also experiment with varying task difficulty, with the hypothesis that stronger environmental pressures will promote the development of more complex language. We find that speakers trained with a ToM listener component have higher accuracies than those trained without in our image referential game setting. We also find that increasing task difficulty in the training process results in more fluent, higher-quality utterances in evaluation. This suggests the utility of incorporating ToM, as well as other insights from child language acquisition, into computational models thereof.

Pareto Invariant Risk Minimization

- Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, James Cheng
- abstract@open-review(Poster): 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.

What Makes Convolutional Models Great on Long Sequence Modeling?

- Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, Debadeepta Dey
- abstract@open-review(Poster): 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.

Editing models with task arithmetic

- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, Ali Farhadi
- abstract@open-review(Poster): 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.

Structured World Representations via Block-Slot Attention

- Gautam Singh, Yeongbin Kim, Sungjin Ahn
- abstract@open-review(Poster): In this paper, we propose a novel object-centric representation, called Block-Slot Representation which, unlike the conventional slot representation, provides concept-level disentanglement within a slot. A block-slot is constructed by composing a set of modular concept representations, called blocks, generated from a memory of concept prototypes. We call this slot construction process Block-Slot Attention. Block-Slot Attention facilitates the emergence of abstract concept blocks within a slot such as color, position, and texture, without any supervision. This brings the benefits of disentanglement into slots and the representation becomes more interpretable. Similar to Slot Attention, this mechanism can be used as a drop-in module in any arbitrary neural architecture. In experiments, we show that our model disentangles object properties significantly better than the previous methods, including complex textured scenes. We also demonstrate the ability to compose novel scenes by composing slots at the block-level.

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
- abstract@open-review(Poster): Large-scale diffusion models have demonstrated remarkable performance on text-to-image synthesis (T2I). Despite their ability to generate high-quality and creative images, users still observe images that do not align well with the text input, especially when involving multiple objects. In this work, we strive to improve the compositional skills of existing large-scale T2I models, specifically more accurate attribute binding and better image compositions. We propose to incorporate language structures with the cross-attention layers based on a recently discovered property of diffusion-based T2I models. Our method is implemented on a state-of-the-art model, Stable Diffusion, and achieves better compositional skills in both qualitative and quantitative results. Our structured cross-attention design is also efficient that requires no additional training samples. 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.

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

- Li-Cheng Lan, Huan Zhang, Cho-Jui Hsieh
- abstract@open-review(Poster): In this paper, we evaluate and improve the generalization performance for rein- forcement learning (RL) agents on the set of “controllable” states, where good policies exist in these states to achieve high rewards. An RL agent that generally masters a task should reach its goal starting from any controllable state of the environment, without memorizing actions specialized for a small set of states. To practically evaluate generalization performance in these states, we propose relay- evaluation, involving starting the test agent from the middle of trajectories of other independently trained, high-reward stranger agents. 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 generalization, we propose a novel method called Self-Trajectory Augmentation (STA), which does not rely on training multiple agents and does not noticeably increase training costs. 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.

CktGNN: Circuit Graph Neural Network for Electronic Design Automation

- Zehao Dong, Weidong Cao, Muhan Zhang, Dacheng Tao, Yixin Chen, Xuan Zhang
- abstract@open-review(Poster): 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 capabilities 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.

Specformer: Spectral Graph Neural Networks Meet Transformers

- Deyu Bo, Chuan Shi, Lele Wang, Renjie Liao
- abstract@open-review(Poster): Spectral graph neural networks 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 graph neural network. 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.

Language Models Can (kind of) Reason: A Systematic Formal Analysis of Chain-of-Thought

- Abulhair Saparov, He He
- abstract@open-review(Poster): 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.

Recursive Time Series Data Augmentation

- Amine Mohamed Aboussalah, Minjae Kwon, Raj G Patel, Cheng Chi, Chi-Guhn Lee
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Understanding the Covariance Structure of Convolutional Filters

- Asher Trockman, Devin Willmott, J Zico Kolter
- abstract@open-review(Poster): 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.

Masked Distillation with Receptive Tokens

- Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, Chang Xu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 Statistical Overfitting

- Qiyang Li, Aviral Kumar, Ilya Kostrikov, Sergey Levine
- abstract@open-review(Poster): 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 statistical overfitting on the temporal-difference (TD) error 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 amount of statistical overfitting. This observation gives us a robust principle for making deep RL efficient: we can hill-climb on a notion of validation temporal-difference error by utilizing any form of regularization techniques from supervised learning. We show that a simple online model selection method that targets the statistical overfitting issue is effective across state-based DMC and Gym tasks.

Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient

- Ming Yin, Mengdi Wang, Yu-Xiang Wang
- abstract@open-review(Poster): \emph{Offline reinforcement learning}, which aims at optimizing sequential decision-making strategies with historical data, has been extensively applied in real-life applications. \emph{State-Of-The-Art} algorithms usually leverage powerful function approximators (\emph{e.g.} neural networks) to alleviate the sample complexity hurdle for better empirical performances. Despite all that, a more systematic understanding of the statistical complexity for function approximation remains lacking. Towards bridging the gap, we take a step by considering offline reinforcement learning with \emph{differentiable function class approximation} (DFA). This function class naturally incorporates a wide range of models with nonlinear/nonconvex structures. Most importantly, we show offline RL with differentiable function approximation is provably efficient by analyzing the \emph{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 improve 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
- abstract@open-review(Poster): 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.

Robust Algorithms on Adaptive Inputs from Bounded Adversaries

- Yeshwanth Cherapanamjeri, Sandeep Silwal, David Woodruff, Fred Zhang, Qiuyi Zhang, Samson Zhou
- abstract@open-review(Poster): We study dynamic algorithms robust to adaptive inputs 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 inputs 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. Together, we provide a unified framework 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. Our general framework has diverse applications in machine learning and data science, such as adaptive distance estimation, kernel density estimation, linear regression, range queries, and point queries. Surprisingly, we show that these novel subroutines for each of these problems can be generally combined with the elegant use of differential privacy to hide the internal randomness of various subroutines, leading to robust algorithms across these different settings. In addition, 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.

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

- Chunhui Zhang, Yijun Tian, Mingxuan Ju, Zheyuan Liu, Yanfang Ye, Nitesh Chawla, Chuxu Zhang
- abstract@open-review(Poster): 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.

On Representing Mixed-Integer Linear Programs by Graph Neural Networks

- Ziang Chen, Jialin Liu, Xinshang Wang, Wotao Yin
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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. Interesting, many of the techniques we explore are complementary to each other to further boost 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.

Simple initialization and parametrization of sinusoidal networks via their kernel bandwidth

- Filipe de Avila Belbute-Peres, J Zico Kolter
- abstract@open-review(Poster): 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 simpler 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.

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
- abstract@open-review(Poster): 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.

Over-Training with Mixup May Hurt Generalization

- Zixuan Liu, Ziqiao Wang, Hongyu Guo, Yongyi Mao
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): The success of large-scale contrastive vision-language pretraining (CLIP) has benefited both visual recognition and multi-modality 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 features can hardly reflect the hierarchy 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
- abstract@open-review(Poster): 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 variety 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 analyze the theoretical properties of our proposed method and demonstrate its ability to rigorously control loss quantiles on several real-world datasets.

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

- Griffin Floto, Stefan Kremer, Mihai Nica
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Learning Achievement Structure for Structured Exploration in Domains with Sparse Reward

- Zihan Zhou, Animesh Garg
- abstract@open-review(Poster): We propose Structured Exploration with Achievements (SEA), a multi-stage reinforcement learning algorithm that learns the environment structure with offline data and uses the learned structure to learn different skills and improve overall exploration with online environment interactions in a particular type of environment that has an internal achievement system. SEA first uses a contrast-based loss function to learn the achievement representations and build an achievement classifier. It then tries to recover the environment achievement structure with a heuristic algorithm. Finally, SEA builds a meta-controller with the recovered structure to learn sub-policies and explore new tasks. While exploration in a procedurally generated environment with high-dimensional input like images is extremely hard for reinforcement learning agents, we demonstrate that SEA is still able to recover the underlying structure and explore new tasks in different domains.

PINTO: Faithful Language Reasoning Using Prompted-Generated Rationales

- PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, Xiang Ren
- abstract@open-review(Poster): 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/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, without any assurance that the 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 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, PINTO leverages the rationales more faithfully than competitive baselines do.

Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods

- Shunta Akiyama, Taiji Suzuki
- abstract@open-review(Poster): 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.

Linearly Mapping from Image to Text Space

- Jack Merullo, Louis Castricato, Carsten Eickhoff, Ellie Pavlick
- abstract@open-review(Poster): 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.

Characterizing intrinsic compositionality in transformers with Tree Projections

- Shikhar Murty, Pratyusha Sharma, Jacob Andreas, Christopher D Manning
- abstract@open-review(Poster): 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.

Augmentation Component Analysis: Modeling Similarity via the Augmentation Overlaps

- Lu Han, Han-Jia Ye, De-Chuan Zhan
- abstract@open-review(Poster): Self-supervised learning aims to learn embeddings with which 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. So the augmentation feature, composed of the distribution of augmentations, can act as the ideal embedding, and similarity over them reveals how much the augmentations of two samples overlap. Without computational burdens to explicitly estimate its value, 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 our method can achieve competitive results against various traditional contrastive learning methods on different benchmarks.

Reproducible Bandits

- Alkis Kalavasis, Grigoris Velegkas, Hossein Esfandiari, Vahab Mirrokni, Andreas Krause, Amin Karbasi
- abstract@open-review(Poster): In this paper, we introduce the notion of reproducible policies in the context of stochastic bandits, one of the canonical problems in interactive learning. A policy in the bandit environment is called reproducible if it pulls, with high probability, the \emph{exact} same sequence of arms in two different and independent executions (i.e., under independent reward realizations). We show that not only do reproducible policies exist, but also they achieve almost the same optimal (non-reproducible) 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 reproducibility parameter is also optimal. Similarly, for stochastic linear bandits (with finitely and infinitely many arms) we develop reproducible policies that achieve the best-known problem-independent regret bounds with an optimal dependency on the reproducibility 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 Bregman Divergences for Distance Learning

- Fred Lu, Edward Raff, Francis Ferraro
- abstract@open-review(Poster): 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.

A Study of Causal Confusion in Preference-Based Reward Learning

- Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, Daniel S. Brown
- abstract@open-review(Poster): 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 aim to study it in the context of reward learning. To study causal confusion, we perform a series of sensitivity and ablation analyses on three 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 the presence of non-causal distractor features, noise in the stated preferences, partial state observability, and larger model capacity can all exacerbate causal confusion. We also identify a set of methods with which to interpret causally confused learned rewards: we observe that optimizing causally confused 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 reward learning to causal confusion, especially in high-dimensional environments---failure to consider even one of many factors (data coverage, state definition, etc.) can quickly 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
- abstract@open-review(Poster): Multi-hop Question Answering over Knowledge Graph~(KGQA) aims to find the answer entities that are multiple hops away from the topic entities 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 firstly retrieves a relatively small subgraph related to the question and then performs the reasoning on the subgraph to accurately find the answer entities. 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 PLM for question-relation semantic matching, and a matching information propagation module to propagate the matching information along the 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.

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

- Shicong Cen, Yuejie Chi, Simon Shaolei Du, Lin Xiao
- abstract@open-review(Poster): 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

- abstract@open-review(Poster): 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 $O(\log n)$ rate is the best possible.

Weighted Clock Logic Point Process

- Ruixuan Yan, Yunshi Wen, Debarun Bhattacharjya, Ronny Luss, Tengfei Ma, Achille Fokoue, Anak Agung Julius
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Low-level sensory and motor signals in the high-dimensional spaces~(e.g., image observations or motor torques) in deep reinforcement learning are complicated to understand or harness for downstream tasks directly. While sensory representations have been widely studied, the representations of actions that form motor skills are yet under exploration. In this work, we find that when a multi-task policy network takes as input states and task embeddings, a space based on the task embeddings emerges to contain meaningful action representations with moderate constraints.Within this space, interpolated or composed embeddings can serve as a high-level interface to instruct the agent to perform meaningful action sequences. Empirical results not only show that the proposed action representations have efficacy for intra-action interpolation and inter-action composition with limited or no learning, but also demonstrate their superior ability in task adaptation to strong baselines in Mujoco locomotion tasks. The evidence elucidates that learning the action representations is a promising direction toward efficient, adaptable, and composable RL, forming the basis of abstract action planning and the understanding of motor signal space. Anonymous project page: https://sites.google.com/view/emergent-action-representation/

Interaction-Based Disentanglement of Entities for Object-Centric World Models

- Akihiro Nakano, Masahiro Suzuki, Yutaka Matsuo
- abstract@open-review(Poster): 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 its ability to learn distinct representations of individual objects and predict their interactions, and it is a focal question how to utilize the learned representations to solve untrained, downstream tasks. 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 yet 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, by excluding features irrelevant in predicting interactions, 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.

Neural Image-based Avatars: Generalizable Radiance Fields for Human Avatar Modeling

- YoungJoong Kwon, Dahun Kim, Duygu Ceylan, Henry Fuchs
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Linear Mode Connectivity of Deep Neural Networks via Permutation Invariance and Renormalization

- Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, Behnam Neyshabur

- abstract@open-review(Poster): In this paper we empirically investigate the conjecture from Entezari et al. (2021) which states that if permutation invariance is taken into account, then there 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.

Diffusion-GAN: Training GANs with Diffusion

- Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, Mingyuan Zhou
- abstract@open-review(Poster): Generative adversarial networks (GANs) are challenging to train stably, and a promising 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.

Mind the Pool: Convolutional Neural Networks Can Overfit Input Size

- Bilal Alsallakh, David Yan, Narine Kokhlikyan, Vivek Miglani, Orion Reblitz-Richardson, Pamela Bhattacharya
- abstract@open-review(Poster): 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.

Reparameterization through Spatial Gradient Scaling

- Alexander Detkov, Mohammad Salameh, Muhammad Fetrat, Jialin Zhang, Robin Luwei, SHANGLING JUI, Di Niu
- abstract@open-review(Poster): 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.

Unsupervised Learning for Combinatorial Optimization Needs Meta Learning

- Haoyu Peter Wang, Pan Li
- abstract@open-review(Poster): A general framework of unsupervised learning for combinatorial optimization (CO) is to train a neural network (NN) whose output gives a problem solution through directly optimizing the CO objective. Albeit with some advantages over traditional solvers, the current framework optimizes 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 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 empirical performance. We observe that even just the initial solution given by our model before fine-tuning can significantly outperform the baselines under various evaluation settings including over the same dataset, cross multiple datasets, and with a shift in the problem scale. The reason we conjecture is that meta-learning-based training may help with finding valleys in the optimization landscape with good local optima for the CO problems that often contain a lot of bad local optima.

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
- abstract@open-review(Poster): A central tenet of Federated learning (FL), which trains models without centralizing user data, is privacy. However, previous work has shown that the gradient updates used in FL can leak user information. While the most industrial uses of FL are for text applications (e.g. keystroke prediction), nearly all attacks on FL privacy have focused on simple image classifiers. 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. This work suggests that FL on text, which has historically been resistant to privacy attacks, is far more vulnerable than previously thought.

Adaptive Optimization in the ∞-Width Limit

- Etai Littwin, Greg Yang
- abstract@open-review(Poster): 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.

Broken Neural Scaling Laws

- Ethan Caballero, Kshitij Gupta, Irina Rish, David Krueger
- abstract@open-review(Poster): We present a smoothly broken power law functional form that accurately models the scaling behaviors (of artificial neural networks) (i.e. how the evaluation metric of interest varies as the amount of compute used for training, number of model parameters, or training dataset size varies) for each task from a very large and diverse set of upstream and downstream (i.e. zero-shot, prompted, and fine-tuned) tasks. These tasks include large-scale vision tasks, large-scale unsupervised language tasks, arithmetic, and reinforcement learning. This functional form yields extrapolations of scaling behavior that often are an order of magnitude more accurate than previous functional forms for modeling the scaling behavior of artificial neural networks. Moreover, this functional form accurately models the non-monotonic transitions present in the scaling behavior of phenomena such as double descent and the delayed, sharp transitions present in the scaling behavior of tasks such as arithmetic.

Avoiding spurious correlations via logit correction

- Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, Carlos Fernandez-Granda
- abstract@open-review(Poster): 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 undesired and poor performance during inference for data lacking such correlations and generalize even worse when more training data present spurious correlations. In this work, we explicitly consider the presence of the potential spurious correlations exist in the majority of training data. Unlike existing approaches which use the ERM model outputs to detect the samples without spurious correlations, and 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, thus the proposed LC could mitigate the negative impacts of spurious correlations in the majority of samples. Our extensive experimental results further reveal that the proposed LC loss outperforms the SoTA solutions on multiple popular benchmarks by a noticeable 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.

Safe Exploration Incurs Nearly No Additional Sample Complexity for Reward-Free RL

- Ruiquan Huang, Jing Yang, Yingbin Liang
- abstract@open-review(Poster): 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.

Diffusion-based Image Translation using disentangled style and content representation

- Gihyun Kwon, Jong Chul Ye
- abstract@open-review(Poster): Modern 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 image translation method using disentangled style and content representation. Specifically, inspired by the slicing Vision Transformer can convert the semantic appearance of a given image into target domain while maintaining the structure of input image, in our method we extract intermediate keys of multihead self attention layer from ViT model and used them as the content preservation loss. More specifically, to preserve the structure information we use the contrastive loss between intermediate keys of the input image and the estimated denoised output during the reverse diffusion sampling. Then, an image guided style transfer is performed by matching the [CLS] token between the denoised diffusion output and target domain, whereas additional CLIP loss is used for the text-driven style transfer.

Implicit Regularization for Group Sparsity

- Jiangyuan Li, Thanh V Nguyen, Chinmay Hegde, Raymond K. W. Wong
- abstract@open-review(Poster): 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.

Large Language Models are Human-Level Prompt Engineers

- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba
- abstract@open-review(Poster): By conditioning on natural language instructions, large language models (LLMs) have 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.

Pruning Deep Neural Networks from a Sparsity Perspective

- Enmao Diao, Ganghua Wang, Jiawei Zhang, Yuhong Yang, Jie Ding, Vahid Tarokh
- abstract@open-review(Poster): 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, the sparsity decreases first when a large model is being effectively regularized and then increases 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 is superior to the state-of-the-art algorithms such as the lottery ticket-based pruning methods, in terms of both compression efficiency and robustness.

Enhancing Meta Learning via Multi-Objective Soft Improvement Functions

- Runsheng Yu, Weiyu Chen, Xinrun Wang, James Kwok
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

GPTQ: Accurate Quantization for Generative Pre-trained Transformers

- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh
- abstract@open-review(Poster): Generative Pre-trained Transformer (GPT) models have set themselves apart by breakthrough performance across complex language modelling tasks, but also by their extremely high computational costs. Specifically, due to memory costs, even inference for large, highly-accurate GPT models may require multiple performant GPUs to execute, which limits their usability. 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 GPTQ, a new one-shot weight quantization method based on approximate second-order information, that is both highly-accurate and highly-efficient. Specifically, GPTQ 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. We show experimentally that these improvements can be leveraged for end-to-end inference speedups over FP16, of around 2x when using high-end GPUs (NVIDIA A100) and 4x when using more cost-effective ones (NVIDIA A6000).

A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution

- Sungyoon Lee, Cheongjae Jang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): We propose a method for 3D shape reconstruction from unoriented point clouds. Our method consists of a novel SE(3)-equivariant coordinate-based network, 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.

Continual Post-Training of Language Models

- Zixuan Ke, Haowei Lin, Yijia Shao, Tatsuya Konishi, Gyuhak Kim, Bing Liu
- abstract@open-review(Poster): Language models (LMs) have been instrumental for the recent rapid advance of natural language processing. Existing research has shown that post-training or adapting an LM using an unlabeled topical/domain corpus can improve the end-task performance in the domain. This paper proposes a novel method to continually post-train an LM with a sequence of unlabeled domain corpora to adapt the LMto these domains to improve their end-task 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 pre-trained LM) and the knowledge from the current full network to achieve knowledge integration. The method not only overcomes catastrophic forgetting, but also achieves knowledge transfer to improve end-task performances compared to post-training each domain separately. 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
- abstract@open-review(Poster): 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.

How Can GANs Learn Hierarchical Generative Models for Real-World Distributions

- Zeyuan Allen-Zhu, Yuanzhi Li
- abstract@open-review(Poster): (this is a theory paper)

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
- abstract@open-review(Poster): 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 hierarchy is not always available, and is often corrupted with missing object descriptions or misaligned bounding box positions. As a result, although using view hierarchy offers some 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 is easily scalable and capable of performing a range of UI modeling tasks. Our experiments show that our model obtains SoTA results on several representative UI tasks and outperforms previous methods that use both screenshots and view hierarchies as input. Furthermore, we explore the multi-task learning and few-shot prompting capacity 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
- abstract@open-review(Poster): 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 three 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.

A Stable and Scalable Method for Solving Initial Value PDEs with Neural Networks

- Marc Anton Finzi, Andres Potapczynski, Matthew Choptuik, Andrew Gordon Wilson
- abstract@open-review(Poster): Unlike conventional grid and mesh based methods for solving PDEs, neural networks have the potential to break the curse of dimensionality, providing approximate solutions to high-dimensional PDEs. While global minimization of the PDE residual over the network parameters works well for boundary value problems, catastrophic forgetting limits its applicability to initial value problems. In an alternative local in time approach, the optimization problem can be converted into an ODE on the network parameters and the solution propagated forward in time; however, we demonstrate that current methods utilizing this idea 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 with 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 high-dimensional PDEs with neural networks.

Heavy-tailed Noise Does Not Explain the Gap Between SGD and Adam, but Sign Descent Might

- Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, Mark Schmidt
- abstract@open-review(Poster): The success of Adam 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 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 stochasticity has heavy tails. This hypothesis suggests that the underlying mechanism causing the gap is a more robust estimator of the gradient. We evaluate this hypothesis by varying the batch size, up to the entire dataset, controlling for stochasticity. We find evidence that stochasticity and heavy-tailed noise are not major factors in the performance gap between SGD and Adam. Rather, SGD does not leverage reductions in noise due to larger batches as well as Adam. This raises the question as to why Adam outperforms SGD in the full-batch setting. Checking simple normalized variants of SGD, we find that the behavior of Adam with increasing batch sizes is most consistent with sign descent.

Building Normalizing Flows with Stochastic Interpolants

- Michael Samuel Albergo, Eric Vanden-Eijnden
- abstract@open-review(Poster): A simple generative model based on a continuous-time normalizing flow between any pair of base and target distributions is proposed. The velocity field of this flow is inferred from the probability current of a time-dependent distribution 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 can be used to estimate the likelihood at any time along the interpolant. The approach is contextualized in its relation to diffusions. In particular, in situations where the base is a Gaussian distribution, we 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. This allows one to map methods based on stochastic differential equations to those of ordinary differential equations, simplifying the mechanics of the model, but capturing equivalent dynamics. Benchmarking on density estimation tasks illustrates that the learned flow can match and surpass maximum likelihood continuous flows at a fraction of the conventional ODE training costs.

Dual Student Networks for Data-Free Model Stealing

- James Beetham, Navid Kardan, Ajmal Saeed Mian, Mubarak Shah
- abstract@open-review(Poster): 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 images 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 with the target model. This push towards disagreeing samples implicitly encourages exploring a more diverse region of 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.

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
- abstract@open-review(Poster): 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.

Equal Improvability: A New Fairness Notion Considering the Long-term Impact

- Ozgur Guldogan, Yuchen Zeng, Jy-yong Sohn, Ramtin Pedarsani, Kangwook Lee
- abstract@open-review(Poster): 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.

Competitive Physics Informed Networks

- Qi Zeng, Yash Kothari, Spencer H Bryngelson, Florian Tobias Schaefer
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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

Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors

- Sizhe Chen, Geng Yuan, Xinwen Cheng, Yifan Gong, Minghai Qin, Yanzhi Wang, Xiaolin Huang
- abstract@open-review(Poster): As data become increasingly vital for deep learning, a company would be very cautious about releasing data. This is because the competitors could use the released data to train high-performance models, thereby posing a tremendous threat to the company's commercial competence. To protect the dataset from unauthorized use for training, imperceptible perturbations crafted with a deep model are added to data so that other deep neural networks trained on it all have poor generalization. In this paper, we propose a self-ensemble protection (SEP) method to take advantage of intermediate checkpoints in a single training process for data protection. Contrary to the popular belief on the similarity of checkpoints, we are surprised to find that their cross-model gradients are close to orthogonal, and thus diverse enough to produce very effective protective perturbations. Besides, we further improve the performance of SEP by developing a novel feature alignment technique to induce feature collapse into the mean of incorrect-class features. Extensive experiments verify the consistent superiority of SEP over 7 state-of-the-art data protection baselines. SEP perturbations on CIFAR-10 with an ℓ_∞ bound as small as $2/255$ could reduce the testing accuracy of a ResNet18 from 94.56% to 14.68%, and the average accuracy reduction from the best-known results is 27.63%. Under the $\ell_\infty=8$ bound, SEP perturbations lead DNNs with 5 architectures to have less than 5.7% / 3.2% / 0.6% accuracy on CIFAR-10 / CIFAR-100 / ImageNet subset.

Effectively Modeling Time Series with Simple Discrete State Spaces

- Michael Zhang, Khaled Kamal Saab, Michael Poli, Tri Dao, Karan Goel, Christopher Re
- abstract@open-review(Poster): 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.

A Time Series is Worth 64 Words: Long-term Forecasting with Transformers

- Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, Jayant Kalagnanam
- abstract@open-review(Poster): 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 and 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
- abstract@open-review(Poster): When learning task-oriented dialogue (TOD) agents, one can naturally utilize reinforcement learning (RL) techniques 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 TOD agents. Specifically, we introduce two generalized objectives for reward-function learning, inspired from the classical learning-to-rank literature. Further, we utilize the learned reward-function to guide the training of the end-to-end TOD agent. With the proposed techniques, we achieve competitive results on the end-to-end response-generation task on the Multiwoz 2.0 dataset.

Supervision Complexity and its Role in Knowledge Distillation

- Hrayr Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, Sanjiv Kumar
- abstract@open-review(Poster): 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.

Transferable Unlearnable Examples

- Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, Jiliang Tang
- abstract@open-review(Poster): With more people publishing their personal data online, unauthorized data usage has become a serious concern. The unlearnable strategies have been introduced to prevent third parties from training on the data without permission. They add perturbations to the users' data before publishing, which aims 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 and datasets. To tackle this issue, we propose a novel unlearnable strategy based on Clustering Separability Discriminant (CSD), which aims to better transfer the unlearnable effects to other training settings and datasets by enhancing the linear separability. Extensive experiments demonstrate the transferability of the proposed unlearnable examples across training settings and datasets.

Random Laplacian Features for Learning with Hyperbolic Space

- Tao Yu, Christopher De Sa
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 quantitatively 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
- abstract@open-review(Poster): 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.

Momentum Stiefel Optimizer, with Applications to Suitably-Orthogonal Attention, and Optimal Transport

- Lingkai Kong, Yuqing Wang, Molei Tao
- abstract@open-review(Poster): 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.

Information-Theoretic Diffusion

- Xianghao Kong, Rob Brekelmans, Greg Ver Steeg
- abstract@open-review(Poster): Denoising diffusion models have spurred significant gains in density modeling and image generation, precipitating an industrial revolution in text-guided AI art generation. Whether interpreted through the lens of variational models or differential equations, diffusion models require many steps of expensive computation to give accurate density estimates. We introduce a new mathematical foundation for diffusion models inspired by classic results in information theory that connects Information with Minimum Mean Square Error estimators, the so-called I-MMSE relations. We generalize the I-MMSE relations to \emph{exactly} relate the data distribution and optimal denoising, leading to an elegant refinement of existing diffusion bounds. This new insight improves density estimation for diffusion models and enables simultaneous modeling of both continuous and discrete probabilities with no additional cost.

SIMPLE: A Gradient Estimator for k-Subset Sampling

- kareem ahmed, Zhe Zeng, Mathias Niepert, Guy Van den Broeck
- abstract@open-review(Poster): 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.

Learning Iterative Neural Optimizers for Image Steganography

- Varsha Kishore, Xiangyu Chen, Kilian Q Weinberger
- abstract@open-review(Poster): Image steganography is the process of concealing secret information in images through imperceptible changes. Recent work has formulated this task as a classical constrained optimization problem. In this paper, we argue that image steganography is inherently performed on the (elusive) manifold of natural images, and propose to train an iterative neural network to perform the optimization steps. In contrast to classical optimization methods like L-BFGS or projected gradient descent, we train a neural network to 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 the previous state-of-the-art encoder-decoder based steganography approaches, it reduces the recovery error rate by multiple orders of magnitude and achieve zero error up to 3 bits per pixel (bpp) without the need for error correcting codes.

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
- abstract@open-review(Poster): 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.

Robust Graph Dictionary Learning

- Weijie Liu, Jiahao Xie, Chao Zhang, Makoto Yamada, Nenggan Zheng, Hui Qian
- abstract@open-review(Poster): 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.

Fundamental limits on the robustness of image classifiers

- Zheng Dai, David Gifford
- abstract@open-review(Poster): 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.

Understanding Influence Functions and Datamodels via Harmonic Analysis

- Nikunj Saunshi, Arushi Gupta, Mark Braverman, Sanjeev Arora
- abstract@open-review(Poster): 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.

TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization

- Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, Shiyu Chang
- abstract@open-review(Poster): 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 training 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.

Information Plane Analysis for Dropout Neural Networks

- Linara Adilova, Bernhard C Geiger, Asja Fischer
- abstract@open-review(Poster): 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: measuring discrete MI does not capture the nature of deterministic neural networks and continuous data distributions, and different approaches for discretization arbitrarily change results. On the other hand, measuring continuous MI for a deterministic network is not mathematically meaningful. In this work we show how the stochasticity induced by dropout layers can be utilized to estimate MI in a theoretically sound manner. We demonstrate in a range of experiments that this approach enables a meaningful information plane analysis for the large class of dropout neural networks that is widely used in practice.

Learning Harmonic Molecular Representations on Riemannian Manifold

- Yiqun Wang, Yuning Shen, Shi Chen, Lihao Wang, Fei YE, Hao Zhou
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 AC and is much less sensitive to entropy-regularization.

Efficiently Controlling Multiple Risks with Pareto Testing

- Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, Tommi S. Jaakkola
- abstract@open-review(Poster): 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.

Characteristic Neural Ordinary Differential Equation

- Xingzi Xu, Ali Hasan, Khalil Elkhalil, Jie Ding, Vahid Tarokh
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 benchmarkpermuted MNIST' than with EWC alone. The benchmark `rotated MNIST' demonstrates how similar tasks recruit similar neurons to the artificial neuronal ensemble.

Learning Language Representations with Logical Inductive Bias

- Jianshu Chen
- abstract@open-review(Poster): 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 devise and compose several neural logic operators into a set of 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 pretrained models (e.g., BERT) 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.

How Does Self-supervised Learning Work? A Representation Learning Perspective

- Yiwen Kou, Zixiang Chen, Yuan Cao, Quanquan Gu
- abstract@open-review(Poster): Self-supervised learning (SSL) is a popular machine learning paradigm that utilizes a large amount of unlabeled data to facilitate the learning from a small number of labeled data. While SSL has achieved great success in different tasks, its theoretical understanding remains largely open. In this paper, we aim to theoretically understand a special kind of SSL approaches based on pre-training and fine-tuning. In particular, the SSL approach we consider first trains a neural network based on the unlabeled data with help of pseudo labelers. Then it fine-tunes the pre-trained network on a small amount of labeled data. We prove that, under certain data and neural network models, SSL 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. Our theoretical result demonstrates a separation between SSL and supervised learning on the same amount of labeled data and sheds light on the essence of representation learning for the success of SSL.

Empowering Graph Representation Learning with Test-Time Graph Transformation

- Wei Jin, Tong Zhao, Jiayuan Ding, Yozen Liu, Jiliang Tang, Neil Shah
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 simple 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.

Interpretations of Domain Adaptations via Layer Variational Analysis

- Huan-Hsin Tseng, Hsin-Yi Lin, Kuo-Hsuan Hung, Yu Tsao
- abstract@open-review(Poster): 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 knowledge adaptation than the gradient descent method.

Denoising Diffusion Samplers

- Francisco Vargas, Will Sussman Grathwohl, Arnaud Doucet
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 heuristics 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.

Interpretable Geometric Deep Learning via Learnable Randomness Injection

- Siqi Miao, Yunan Luo, Mia Liu, Pan Li
- abstract@open-review(Poster): Point cloud data is ubiquitous in scientific fields. Recently, geometric deep learning (GDL) has been widely applied to solve prediction tasks with point cloud data. However, GDL models are often complicated and hardly interpretable, which poses concerns to scientists when deploying these models in scientific analysis and experiments. This work proposes a general mechanism based on \learnable randomness injection (LRI) that allows building inherently interpretable models with general GDL backbones. Once being trained, LRI-induced models may detect the points within the point cloud data, which carry information indicative to the prediction labels. Such indicative information may be reflected by either the existence of these points in the data or the geometric locations of these points. We also propose four scientific datasets in the domains of high energy physics and biochemistry to evaluate LRI. 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-induced models are also more robust to the distribution shifts between training and test scenarios.

GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure

- Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, Mihaela van der Schaar
- abstract@open-review(Poster): 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.

Progressive Prompts: Continual Learning for Language Models without Forgetting

- Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, Amjad Almahairi
- abstract@open-review(Poster): We introduce Progressive Prompts - a simple and efficient approach for continual learning in language models. Our method does not require data replay, and alleviates catastrophic forgetting without using 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.

Deep Learning From Crowdsourced Labels: Coupled Cross-Entropy Minimization, Identifiability, and Regularization

- Shahana Ibrahim, Tri Nguyen, Xiao Fu
- abstract@open-review(Poster): 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 identifiability 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.

Projective Proximal Gradient Descent for Nonconvex Nonsmooth Optimization: Fast Convergence Without Kurdyka-Lojasiewicz (KL) Property

- Yingzhen Yang, Ping Li
- abstract@open-review(Poster): 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 nd 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-Lojasiewicz (KL) property, we provide a new theoretical analysis showing that PPGD achieves optimal convergence rate on a class of nonconvex and nonsmooth problems under mild assumptions, which is the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continuous gradient. Experimental results demonstrate the effectiveness of the PPGD.

First Steps Toward Understanding the Extrapolation of Nonlinear Models to Unseen Domains

- Kefan Dong, Tengyu Ma
- abstract@open-review(Poster): 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.

Variable Compositionality Reliably Emerges in Neural Networks

- Henry Conklin, Kenny Smith
- abstract@open-review(Poster): Human languages are pervasively compositional: because the meaning of an expression is composed from the meaning of its parts, we can productively leverage our prior experience in order to communicate about novel meanings. Work looking at the languages that emerge when neural networks solve communicative tasks has shown that networks regularly develop languages that allow them to communicate and generalize to novel examples; surprisingly, that work has struggled to show that compositional systems reliably develop, 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, just with variation. We introduce a variation-based framework for interpreting the mappings produced by neural networks in emergent communication games and find that they reliably exhibit straight-forward compositional structure, with a degree of natural language-like variation that obscures their compositionality under measures used in previous work. We show that early in training measures of variation correlate with generalization performance, but that this effect goes away later in training as the languages become regular enough to compositionally generalize. In an effort to decrease the variability of the 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 novel methodology for aligning sequences using a natural language processing (NLP) approach. BetaAlign accounts for the possible variability of the evolutionary process among different datasets by using an ensemble of transformers, each trained on millions of samples generated from a different evolutionary model. Our approach leads to outstanding alignment accuracy, often outperforming commonly used methods, such as MAFFT, DIALIGN, ClustalW, T-Coffee, and MUSCLE.

A Mixture-of-Expert Approach to RL-based Dialogue Management

- Yinlam Chow, Azamat Tulepbergenov, Ofir Nachum, Dhawal Gupta, Moonkyung Ryu, Mohammad Ghavamzadeh, Craig Boutilier
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 denoising 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 tasks such as online, few-shot, and continual learning, where brains are known to excel. Despite this promising empirical performance, 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.

The Onset of Variance-Limited Behavior for Networks in the Lazy and Rich Regimes

- Alexander Atanasov, Blake Bordelon, Sabarish Sainathan, Cengiz Pehlevan
- abstract@open-review(Poster): 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, at a critical sample size P^*, *the finite-width network generalization begins to worsen compared to the infinite width performance. In this work, we empirically study the transition from the infinite width behavior to this 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 going as $P^* \sim \sqrt{N}$* for polynomial regression with ReLU networks. We discuss the source of this finite size behavior based on the variance of the NN's final neural tangent kernel (NTK). We then show how this transition can be pushed to larger P by enhancing feature learning or by ensemble averaging the network. 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
- abstract@open-review(Poster): 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. On the AI2-THOR Rearrangement Challenge, our method improves on current state-of-the-art end-to-end reinforcement learning-based methods that learn visual rearrangement policies from 0.53\% correct rearrangement to 16.56\%, using only 2.7\% as many samples from the environment.

Progressive Mix-Up for Few-Shot Supervised Multi-Source Domain Transfer

- Ronghang Zhu, Ronghang Zhu, Xiang Yu, Sheng Li
- abstract@open-review(Poster): 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.

Neural Compositional Rule Learning for Knowledge Graph Reasoning

- Kewei Cheng, Nesreen Ahmed, Yizhou Sun
- abstract@open-review(Poster): Learning logic rules is critical to improve reasoning in knowledge graphs. This is due to their ability to provide logical 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 logic rules, the majority of these methods are either restricted by their computational complexity to manage the large search space, or are particularly designed for inductive relational reasoning task. In this paper, we propose an end-to-end neural model for learning compositional logic rules called NCRL. NCRL treats logic rules as a hierarchical tree, and breaks the rule body into small atomic compositions during inference. 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 across multiple tasks. Specifically, we show that NCRL is scalable and yields state-of-the-art results for link prediction on large-scale KGs. Moreover, we test NCRL for systematic generalization by learning to reason on small-scale observed graphs and evaluating on larger ones.

Efficient approximation of neural population structure and correlations with probabilistic circuits

- Koosha Khalvati, Samantha Johnson, Stefan Mihalas, Michael A Buice
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, robustness, biologically-inspired filtering mechanisms, and representational straightness in response to time-varying input; we identify curvature 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. Finally, (3) biologically-inspired elements increase straightness in the early stages of a network, but do not guarantee increased straightness in downstream layers of CNNs. Our results suggest that the ability of a model to straighten is a useful and easily computed measure of representational robustness and stability, as well as a marker of similarity between human and machine representations.

Is Forgetting Less a Good Inductive Bias for Forward Transfer?

- Jiefeng Chen, Arslan Chaudhry, Timothy Nguyen, Dilan Gorur
- abstract@open-review(Poster): 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 restricted updates on the task by the continual learner to preserve 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.

Learning Structured Representations by Embedding Class Hierarchy

- Siqi Zeng, Remi Tachet des Combes, Han Zhao
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 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 biological utility, we found that RS was the best network for explaining neurons in primate V1 providing a parsimonious, bio-plausible mechanism (noise in the sensory input stages) for generating representations in the early visual cortex. Finally, we compared the geometric properties of frame representations in these networks to better understand how they produced brain-like representations. 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 more brain-like straightened movie representations that predict neural behavior.

FunkNN: Neural Interpolation for Functional Generation

- AmirEhsan Khorashadizadeh, Anadi Chaman, Valentin Debarnot, Ivan Dokmanić
- abstract@open-review(Poster): 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.

Label Propagation with Weak Supervision

- Rattana Pukdee, Dylan Sam, Pradeep Kumar Ravikumar, Nina Balcan
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): There has been growing interest in automatically predicting missing type annotations in programs written in Python and Javascript. Prior methods have achieved impressive accuracy when predicting the most common types, but they often perform poorly on rare or complex types (or do not support them at all). In this paper, we present a new type inference method that treats type prediction as a code completion task by leveraging CodeT5, a state-of-the-art seq2seq pre-trained language model for code. Our method uses static analysis to construct context for each code element whose type is to be predicted by the model. We also propose a sequential decoding scheme that incorporates previous type predictions in the model's input context, allowing information exchange between related code elements. Our evaluation shows that our 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.

AGRO: Adversarial discovery of error-prone Groups for Robust Optimization

- Bhargavi Paranjape, Pradeep Dasigi, Vivek Srikumar, Luke Zettlemoyer, Hannaneh Hajishirzi
- abstract@open-review(Poster): 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.

LogicDP: Creating Labels for Graph Data via Inductive Logic Programming

- Yuan Yang, Faramarz Fekri, James Clayton Kerce, Ali Payani
- abstract@open-review(Poster): 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 large 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.

Revisiting Curiosity for Exploration in Procedurally Generated Environments

- Kaixin Wang, Kuangqi Zhou, Bingyi Kang, Jiashi Feng, Shuicheng YAN
- abstract@open-review(Poster): Exploration under sparse rewards remains a key challenge in deep reinforcement learning. Recently, studying exploration in procedurally-generated environments draws increasing attention. Existing works generally combine lifelong curiosity and episodic curiosity as the intrinsic reward to encourage exploration. Though various lifelong and episodic curiosities have been proposed, the individual contributions of the two kinds of curiosities to improving exploration are barely investigated. To bridge this gap, we disentangle these two parts and conduct extensive ablative experiments. We consider lifelong and episodic curiosities 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 curiosity can match or surpass prior state-of-the-art methods. On the other hand, only using lifelong curiosity can hardly make progress in exploration. This demonstrates that episodic curiosity is more crucial than lifelong curiosity in boosting exploration. Moreover, we find through experimental analysis that the learned lifelong curiosity does not accurately reflect the novelty of states, which explains why it does not help much in improving exploration.

Transformer-based World Models Are Happy With 100k Interactions

- Jan Robine, Marc Höftmann, Tobias Uelwer, Stefan Harmeling
- abstract@open-review(Poster): 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.

Can Neural Networks Learn Implicit Logic from Physical Reasoning?

- Aaron Traylor, Roman Feiman, Ellie Pavlick
- abstract@open-review(Poster): 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 otherwise good representations of the objects' physical dynamics and

constraints, 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.

ESCHER: Eschewing Importance Sampling in Games by Computing a History Value Function to Estimate Regret

- Stephen Marcus McAleer, Gabriele Farina, Marc Lanctot, Tuomas Sandholm
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Towards Understanding GD with Hard and Conjugate Pseudo-labels for Test-Time Adaptation

- Jun-Kun Wang, Andre Wibisono
- abstract@open-review(Poster): 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 a solution that minimizes the testing 0-1 loss under a Gaussian model, 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.

A VAE for Transformers with Nonparametric Variational Information Bottleneck

- James Henderson, Fabio James Fehr
- abstract@open-review(Poster): We propose a Variational AutoEncoder (VAE) for Transformers by developing a Variational Information Bottleneck (VIB) regulariser for Transformer embeddings. We formalise the embedding space of Transformer encoders as mixture distributions, and use Bayesian nonparametrics to develop a Nonparametric VIB (NVIB) for such attention-based representations. The variable number of mixture components supported by nonparametric methods captures the variable number of vectors supported by attention, and exchangeable distributions from nonparametric methods capture the permutation invariance of attention. We then propose our Transformer VAE (NVAE) using NVIB to regularise the information passing from the Transformer encoder to the Transformer decoder through cross-attention. 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.

On The Specialization of Neural Modules

- Devon Jarvis, Richard Klein, Benjamin Rosman, Andrew M Saxe
- abstract@open-review(Poster): 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 linear 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
- abstract@open-review(Poster): 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.

Using Both Demonstrations and Language Instructions to Efficiently Learn Robotic Tasks

- Albert Yu, Ray Mooney
- abstract@open-review(Poster): 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.

FIGARO: Controllable Music Generation using Learned and Expert Features

- Dimitri von Rütte, Luca Biggio, Yannic Kilcher, Thomas Hofmann
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 show 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.

Recitation-Augmented Language Models

- Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, Denny Zhou
- abstract@open-review(Poster): 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).

KwikBucks: Correlation Clustering with Cheap-Weak and Expensive-Strong Signals

- Sandeep Silwal, Sara Ahmadian, Andrew Nystrom, Andrew McCallum, Deepak Ramachandran, Seyed Mehran Kazemi
- abstract@open-review(Poster): 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.

Reward Design with Language Models

- Minae Kwon, Sang Michael Xie, Kalesha Bullard, Dorsa Sadigh
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 this 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.

A Statistical Framework for Personalized Federated Learning and Estimation: Theory, Algorithms, and Privacy

- Kaan Ozkara, Antonious M. Girgis, Deepesh Data, Suhas Diggavi
- abstract@open-review(Poster): 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 generative framework that could potentially unify several different algorithms as well as suggest new algorithms. We apply our generative framework to personalized estimation, and connect it to the classical empirical Bayes' methodology. We develop private personalized estimation under this framework. We then use our generative framework to propose new personalized learning algorithms, including AdaPeD based on a Knowledge Distillation, 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.

Subsampling in Large Graphs Using Ricci Curvature

- Shushan Wu, Huimin Cheng, Jiazhang Cai, Ping Ma, Wenxuan Zhong
- abstract@open-review(Poster): 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.

Conservative Bayesian Model-Based Value Expansion for Offline Policy Optimization

- Jihwan Jeong, Xiaoyu Wang, Michael Gimelfarb, Hyunwoo Kim, Baher abdulhai, Scott Sanner
- abstract@open-review(Poster): 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.

Scaling up and Stabilizing Differentiable Planning with Implicit Differentiation

- Linfeng Zhao, Huazhe Xu, Lawson L.S. Wong
- abstract@open-review(Poster): 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.

Score-based Continuous-time Discrete Diffusion Models

- Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, Hanjun Dai
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 requires considerable online interaction data that are costly and unsafe to collect in the real world. Furthermore, when they are applied to the frame dropping scenarios, they perform unsatisfactorily even with moderate drop rates. To devise a robust and deployable algorithm, 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 timespan 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.

Adversarial Imitation Learning with Preferences

- Aleksandar Taranovic, Andras Gabor Kupcsik, Niklas Freymuth, Gerhard Neumann
- abstract@open-review(Poster): 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.

Is Model Ensemble Necessary? Model-based RL via a Single Model with Lipschitz Regularized Value Function

- Ruijie Zheng, Xiyao Wang, Huazhe Xu, Furong Huang
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 the challenge of building faithful generative models for many-to-many datasets. We then present a novel, scalable generation framework based on 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 Low Dimensional State Spaces with Overparameterized Recurrent Neural Networks

- Edo Cohen-Karlik, Itamar Menuhin-Gruman, Nadav Cohen, Raja Giryes, Amir Globerson
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 specialized 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.

Why (and When) does Local SGD Generalize Better than SGD?

- Xinran Gu, Kaifeng Lyu, Longbo Huang, Sanjeev Arora
- abstract@open-review(Poster): 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 with a small learning rate, after approaching the manifold of 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 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
- abstract@open-review(Poster): 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 risk-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.

Analogical Networks for Memory-Modulated 3D Parsing

- Nikolaos Gkanatsios, Mayank Singh, Zhaoyuan Fang, Shubham Tulsiani, Katerina Fragkiadaki
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Integrating Symmetry into Differentiable Planning with Steerable Convolutions

- Linfeng Zhao, Xupeng Zhu, Lingzhi Kong, Robin Walters, Lawson L.S. Wong
- abstract@open-review(Poster): We study how group symmetry helps improve data efficiency and generalization for end-to-end differentiable planning algorithms, when symmetry appears in decision-making tasks. Motivated by equivariant convolution networks, we treat the path planning problem as \textit{signals} over grids. We show that value iteration in this case is a \textit{linear equivariant operator}, which is a (steerable) \textit{convolution}. This extends Value Iteration Networks (VINs) on using convolutional networks for path planning with additional \textit{rotation} and \textit{reflection} symmetry. Our implementation is based on VINs and uses steerable convolution networks to incorporate symmetry. The experiments are performed on four tasks: 2D navigation, visual navigation, 2 degrees of freedom (2DOFs) configuration space and workspace manipulation. % in configuration space or workspace. Our symmetric planning algorithms improve training efficiency and generalization by large margins compared to non-equivariant counterparts, VIN and GPPN.

Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning

- Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, Cheng Zhang
- abstract@open-review(Poster): 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.

$O(T^{-1})$ Convergence of Optimistic-Follow-the-Regularized-Leader in Two-Player Zero-Sum Markov Games

- Yuepeng Yang, Cong Ma
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 variation due to group actions. Here, we demonstrate that BNNs are able to simultaneously learn groups, their irreducible representations, and corresponding complete invariant maps purely from symmetries implicit in data. Further, we demonstrate that the completeness property endows these networks with strong adversarial robustness. This work establishes Bispectral Neural Networks as a powerful computational primitive for robust invariant representation learning.

Beyond Lipschitz: Sharp Generalization and Excess Risk Bounds for Full-Batch GD

- Konstantinos Nikolakakis, Farzin Haddadpour, Amin Karbasi, Dionysios Kalogerias
- abstract@open-review(Poster): 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).

Hyper-Decision Transformer for Efficient Online Policy Adaptation

- Mengdi Xu, Yuchen Lu, Yikang Shen, Shun Zhang, Ding Zhao, Chuang Gan

- abstract@open-review(Poster): 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.

Solving Continuous Control via Q-learning

- Tim Seyde, Peter Werner, Wilko Schwarting, Igor Gilitschenski, Martin Riedmiller, Daniela Rus, Markus Wulfmeier
- abstract@open-review(Poster): While there has been substantial success in applying actor-critic methods to continuous control, simpler critic-only methods such as Q-learning often remain intractable in the associated high-dimensional action spaces. However, most actor-critic methods come at the cost of added complexity: heuristics for stabilisation, compute requirements as well as wider hyperparameter search spaces. We show that these issues can be largely alleviated via Q-learning by combining action discretization with value decomposition, framing single-agent control as cooperative multi-agent reinforcement learning (MARL). With bang-bang actions, performance of this critic-only approach matches 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 surprisingly strong performance across a wide 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
- abstract@open-review(Poster): 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.

Personalized Reward Learning with Interaction-Grounded Learning (IGL)

- Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, Cheng Tan
- abstract@open-review(Poster): 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 a fixed combination of implicit feedback signals across all users. However, this approach disregards a growing body of work 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 taking 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.

Towards convergence to Nash equilibria in two-team zero-sum games

- Fivos Kalogiannis, Ioannis Panageas, Emmanouil-Vasileios Vlatakis-Gkaragkounis
- abstract@open-review(Poster): 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.

Meta-Learning Black-Box Optimization via Black-Box Optimization

- Robert Tjarko Lange, Tom Schaul, Yutian Chen, Tom Zahavy, Valentin Dalibard, Chris Lu, Satinder Singh, Sebastian Flennerhag
- abstract@open-review(Poster): 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.

DensePure: Understanding Diffusion Models towards Adversarial Robustness

- Zhongzhu Chen, Kun Jin, Chaowei Xiao, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, Dawn Song
- abstract@open-review(Poster): 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.

Grounding Graph Network Simulators using Physical Sensor Observations

- Jonas Linkerhägner, Niklas Freymuth, Paul Maria Scheikl, Franziska Mathis-Ullrich, Gerhard Neumann
- abstract@open-review(Poster): 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 generated 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.

Where to Diffuse, How to Diffuse and How to get back: Learning in Multivariate Diffusions

- Raghav Singhal, Mark Goldstein, Rajesh Ranganath
- abstract@open-review(Poster): Diffusion-based generative models (DBGMs) perturb data to a target noise distribution and reverse this inference process to generate samples. The choice of inference diffusion affects both likelihoods and sample quality as it is tied to the generative model. Recent work in DBGMs has applied the principle of improving generative models with the use of auxiliary variables, leading to improved sample quality. While there are many such multivariate diffusions to explore, each new one requires significant model-specific analysis, hindering rapid prototyping and evaluation. In this work, we study linear Multivariate Diffusion Models (MDMs). First, for any number of auxiliary variables, we provide a recipe for maximizing a lower-bound on the MDM likelihood, without requiring any model-specific analysis. Next, we 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 the set of linear diffusions. To demonstrate these ideas, we introduce two new specific diffusions as well as learn a diffusion process on the MNIST and CIFAR10 datasets. We achieve improved bits-per-dim bounds using the new diffusion, compared to the existing likelihood-trained VPSDE. We additionally connect the existing CLD objective to the likelihood lower-bound.

Contrastive Corpus Attribution for Explaining Representations

- Chris Lin, Hugh Chen, Chanwoo Kim, Su-In Lee
- abstract@open-review(Poster): 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).

Spatio-temporal point processes with deep non-stationary kernels

- Zheng Dong, Xiuyuan Cheng, Yao Xie
- abstract@open-review(Poster): Deep neural networks, especially recurrent neural network (RNN) models, have become a popular tool for analyzing point process data. Despite the powerful expressiveness and memorizing ability of RNN models, they may not successfully model sophisticated non-stationary dependencies among data due to the recurrent structure. Meanwhile, another type of deep model for point process data was recently proposed, which represents the influence kernel rather than the intensity function by neural networks. This paper develops a deep non-stationary influence kernel for spatio-temporal point processes with a novel parameterization that enables us to well approximate complicated kernels in a low-rank form. A log-barrier penalty is introduced during network optimization to maintain the non-negativity of conditional intensity. Our new method can also be extended to model high-dimensional marks, and we demonstrate outstanding performance gain on real police text data. The new approach significantly reduces the model and computational complexities, and the benefits of kernel recovery and event prediction are demonstrated using synthetic and real point process data.

Federated Learning from Small Datasets

- Michael Kamp, Jonas Fischer, Jilles Vreeken
- abstract@open-review(Poster): 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, we so achieve effective learning even if the local datasets are extremely small, while retaining the privacy benefits of federated learning.

Relative Behavioral Attributes: Filling the Gap between Symbolic Goal Specification and Reward Learning from Human Preferences

- Lin Guan, Karthik Valmeekam, Subbarao Kambhampati
- abstract@open-review(Poster): Generating complex behaviors that satisfy the preferences of non-expert users is a crucial requirement on AI agents. Interactive reward learning from trajectory comparisons 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.

Scalable Batch-Mode Deep Bayesian Active Learning via Equivalence Class Annealing

- Renyu Zhang, Aly A Khan, Robert L. Grossman, Yuxin Chen
- abstract@open-review(Poster): 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.

Semi Parametric Inducing Point Networks

- Richa Rastogi, Yair Schiff, Alon Hacohen, Zhaozhi Li, Ian Lee, Yuntian Deng, Mert R. Sabuncu, Volodymyr Kuleshov
- abstract@open-review(Poster): 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 and improves accuracy across a range of meta-learning tasks and improves state-of-the-art performance on an important practical problem, genotype imputation.

DAG Learning via Sparse Relaxations

- Valentina Zantedeschi, Luca Franceschi, Jean Kaddour, Matt Kusner, Vlad Niculae
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction

- Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, Jianzhu Ma
- abstract@open-review(Poster): Rich data and powerful machine learning models allow us to design drugs for a specific protein target *in silico*. 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.

How gradient estimator variance and bias impact learning in neural networks

- Arna Ghosh, Yuhan Helena Liu, Guillaume Lajoie, Konrad Kording, Blake Aaron Richards
- abstract@open-review(Poster): 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 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.

Evaluating Representations with Readout Model Switching

- Yazhe Li, Jorg Bornschein, Marcus Hutter
- abstract@open-review(Poster): 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 the model complexity, as well as the data efficiency into account. As a result, the most appropriate model for the specific task and representation will 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. Finally, we discuss important properties revealed by these evaluations such as model scaling, preferred readout model, and data efficiency.

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

- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 state-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.

Planning with Language Models through Iterative Energy Minimization

- Hongyi Chen, Yilun Du, Yiye Chen, Patricio A. Vela, Joshua B. Tenenbaum
- abstract@open-review(Poster): 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.

The Union of Manifolds Hypothesis

- Bradley CA Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C Cresswell, Gabriel Loaiza-Ganem
- abstract@open-review(Poster): 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 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 put forth the union of manifolds hypothesis, which states that data lies on a disjoint union of manifolds of varying intrinsic dimensionality. 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 impact of the union of manifolds hypothesis in deep learning, both supervised and unsupervised, showing that designing models with an inductive bias towards this structure improves performance across classification and generative modelling tasks.

Error Sensitivity Modulation based Experience Replay: Mitigating Abrupt Representation Drift in Continual Learning

- Fahad Sarfraz, Elahe Arani, Bahram Zonooz
- abstract@open-review(Poster): 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 error-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 for modulating the error sensitivity in a dual-memory rehearsal-based system. Concretely, it maintains a memory of past errors and utilizes 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 streams.

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
- abstract@open-review(Poster): Out of distribution (OoD) generalization has received considerable interest in recent 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.

ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond

- Xiaojun Guo, Yifei Wang, Tianqi Du, Yisen Wang
- abstract@open-review(Poster): 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

- abstract@open-review(Poster): We study first-order methods for constrained min-max optimization. Existing methods either requires two gradient calls or two projections in each iteration, which may be costly in applications. In this paper, we first show that the Optimistic Gradient (OG) method, a single-call single-projection} algorithm, has $O(\frac{1}{\sqrt{T}})$ 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 Accelerated Reflected Gradient (ARG) method that achieves the optimal $O(\frac{1}{T})$ 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 Reflected Gradient (RG) method, another 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)

Distributed Extra-gradient with Optimal Complexity and Communication Guarantees

- Ali Ramezani-Kebrya, Kimon Antonakopoulos, Igor Krawczuk, Justin Deschenaux, Volkan Cevher
- abstract@open-review(Poster): 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.

Performance Bounds for Model and Policy Transfer in Hidden-parameter MDPs

- Haotian Fu, Jiayu Yao, Omer Gottesman, Finale Doshi-Velez, George Konidaris
- abstract@open-review(Poster): 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.

Compositional Task Generalization with Discovered Successor Feature Modules

- Wilka Torrico Carvalho, Angelos Filos, Richard Lewis, Honglak Lee, Satinder Singh
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Effective passive membership inference attacks in federated learning against overparameterized models

- Jiacheng Li, Ninghui Li, Bruno Ribeiro
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 standard 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) and high-value verticals such as biomedicine (e.g., GENIA, NCBI, BC5CDR, JNLPBA).

Taking a Step Back with KCal: Multi-Class Kernel-Based Calibration for Deep Neural Networks

- Zhen Lin, Shubhendu Trivedi, Jimeng Sun
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Differentially Private Adaptive Optimization with Delayed Preconditioners

- Tian Li, Manzil Zaheer, Ken Liu, Sashank J. Reddi, Hugh Brendan McMahan, Virginia Smith
- abstract@open-review(Poster): 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 preconditioners 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.

Long Range Language Modeling via Gated State Spaces

- Harsh Mehta, Ankit Gupta, Ashok Cutkosky, Behnam Neyshabur
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Investigating Multi-task Pretraining and Generalization in Reinforcement Learning

- Adrien Ali Taiga, Rishabh Agarwal, Jesse Farebrother, Aaron Courville, Marc G Bellemare
- abstract@open-review(Poster): Deep reinforcement learning (RL) has achieved remarkable successes in complex single-task settings. However, learning policies that can perform multiple tasks and leverage prior experience to learn faster remains challenging. Despite previous attempts to improve on these areas, our understanding of multi-task training and generalization in reinforcement learning remains limited. In this work we propose to investigate the generalization capabilities of a popular actor-critic method, IMPALA. We build on previous work that has advocated for the use of modes and difficulties of Atari 2600 games as a benchmark for transfer learning in reinforcement learning. We do so by pretraining an agent on multiple flavours of the same game before finetuning on the remaining unseen ones. 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 of a game are able to generalize better. Surprisingly we observe that this advantage can be more pronounced after finetuning for 200M environment frames than when doing zero-shot transfer. This highlights the importance of the learned representation and that performance after finetuning might more appropriate to evaluate generalization in reinforcement learning. 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 crucial 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

CrAM: A Compression-Aware Minimizer

- Alexandra Peste, Adrian Vladu, Eldar Kurtic, Christoph H Lampert, Dan Alistarh
- abstract@open-review(Poster): 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: for instance, on the ImageNet task, we can prune models in one-shot to 70-80% sparsity with reasonable ($\leq$ 1%) accuracy loss, which is competitive with gradual compression methods. Additionally, we show that CrAM produces sparse models which perform well for transfer learning, and that it also works for semi-structured pruning patterns supported by GPU hardware.

Understanding Train-Validation Split in Meta-Learning with Neural Networks

- Xinzhe Zuo, Zixiang Chen, Huaxiu Yao, Yuan Cao, Quanquan Gu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): There is a growing interest in the machine learning community in developing predictive algorithms that are ``interpretable by design". Towards this end, recent work proposes to make interpretable decisions by sequentially asking interpretable queries about data until a prediction can be made with high confidence based on the answers obtained (the history). To promote 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 for learning generative models. V-IP is based on finding a query selection strategy and a classifier that minimizes the expected cross-entropy between true and predicted labels. We then demonstrate 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 modelling approach.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, unlike in other domains, 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.

Logical Message Passing Networks with One-hop Inference on Atomic Formulas

- Zihao Wang, Yangqiu Song, Ginny Wong, Simon See
- abstract@open-review(Poster): 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 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 in 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 representation 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.

Noise-Robust De-Duplication at Scale

- Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, Melissa Dell
- abstract@open-review(Poster): 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. The public release of our NEWS-COPY de-duplication dataset will facilitate further research and applications.

Few-shot Backdoor Attacks via Neural Tangent Kernels

- Jonathan Hayase, Sewoong Oh
- abstract@open-review(Poster): 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.

Hyperparameter Optimization through Neural Network Partitioning

- Bruno Kacper Mlodozeniec, Matthias Reisser, Christos Louizos
- abstract@open-review(Poster): 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.

Symmetries, Flat Minima and the Conserved Quantities of Gradient Flow

- Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, Nima Dehmamy
- abstract@open-review(Poster): Empirical studies have revealed that many minima are connected and reside in low-loss valleys in the loss landscape of deep networks. Ensemble models sam- pling different parts of a low-loss valley have reached SOTA performance. Yet, little is known about the theoretical origin of these low-loss valleys. We present a general framework for finding continuous symmetries in the parameter space, which give rise to the low-loss valleys. Importantly, we introduce a novel set of nonlinear, data-dependent symmetries for neural networks. These symmetries can transform a trained model such that it performs similarly on new samples. 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. We also find the nonlinear action to be viable for ensemble building to improve robustness under certain adversarial attacks.

Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees

- Swarnadeep Saha, Shiyue Zhang, Peter Hase, Mohit Bansal
- abstract@open-review(Poster): 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 supervision, 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 hu- man 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
- abstract@open-review(Poster): 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 complete 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.

Architectural optimization over subgroups of equivariant neural networks

- Kaitlin Maile, Dennis George Wilson, Patrick Forré
- abstract@open-review(Poster): 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.

Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time

- Jun-Kun Wang, Andre Wibisono
- abstract@open-review(Poster): 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 design 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.

Order Matters: Agent-by-agent Policy Optimization

- Xihuai Wang, Zheng Tian, Ziyu Wan, Ying Wen, Jun Wang, Weinan Zhang
- abstract@open-review(Poster): 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.

On the Convergence of AdaGrad on $\mathbb{R}^d$: Beyond Convexity, Non-Asymptotic Rate and Acceleration

- Zijian Liu, Ta Duy Nguyen, Alina Ene, Huy Nguyen
- abstract@open-review(Poster): 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 which 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.

SP2 : A Second Order Stochastic Polyak Method

- Shuang Li, William Joseph Swartworth, Martin Takáč, Deanna Needell, Robert M. Gower
- abstract@open-review(Poster): 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.

Making Better Decision by Directly Planning in Continuous Control

- Jinhua Zhu, Yue Wang, Lijun Wu, Tao Qin, Wengang Zhou, Tie-Yan Liu, Houqiang Li
- abstract@open-review(Poster): By properly utilizing the learned environment model, model-based reinforcement learning 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.

HiT-MDP: Learning the SMDP option framework on MDPs with Hidden Temporal Variables

- Chang Li, Dongjin Song, Dacheng Tao
- abstract@open-review(Poster): 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 a novel Markov Decision Process (MDP), the Hidden Temporal MDP (HiT-MDP), and prove that the option-induced HiT-MDP is homomorphic equivalent to the option-induced SMDP. We also derive a sample efficient structured variational inference-based algorithm which leads to a novel stable option discovering method under the maximum-entropy reinforcement learning framework. Extensive experiments on challenging \textit{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.

(Certified!!) Adversarial Robustness for Free!

- Nicholas Carlini, J Zico Kolter, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun
- abstract@open-review(Poster): 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.

Heterogeneous Neuronal and Synaptic Dynamics for Spike-Efficient Unsupervised Learning: Theory and Design Principles

- Biswadeep Chakraborty, Saibal Mukhopadhyay
- abstract@open-review(Poster): 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 the integration/relaxation dynamics of neurons 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 \textbf{h}eterogeneous RSNN (HRSNN) 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 optimized HRSNN increases performance and reduces spiking activity compared to a ho\textbf{m}ogeneous RSNN (MRSNN).

MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises

- Emanuele Palumbo, Imant Daunhawer, Julia E Vogt
- abstract@open-review(Poster): 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.

In-Situ Text-Only Adaptation of Speech Models with Low-Overhead Speech Imputations

- Ashish Mittal, Sunita Sarawagi, Preethi Jyothi
- abstract@open-review(Poster): 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.

Scaling Laws For Deep Learning Based Image Reconstruction

- Tobit Klug, Reinhard Heckel
- abstract@open-review(Poster): 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 optimally 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.

Meta Learning to Bridge Vision and Language Models for Multimodal Few-Shot Learning

- Ivona Najdenkoska, Xiantong Zhen, Marcel Worring
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Improved Learning-augmented Algorithms for k-means and k-medians Clustering

- Thy Dinh Nguyen, Anamay Chaturvedi, Huy Nguyen
- abstract@open-review(Poster): 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 $(1+O(\alpha))\OPT$, while requiring essentially just $O(md \log^3 m/\alpha)$ runtime.

Neural Implicit Shape Editing using Boundary Sensitivity

- Arturs Berzins, Moritz Ibing, Leif Kobbelt
- abstract@open-review(Poster): 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 \emph{boundary sensitivity} to express how perturbations in parameters move the shape boundary. This allows to interpret the effect of each learnable parameter and study achievable deformations. With this, we perform \emph{geometric editing}: finding a parameter update which best approximates a globally prescribed deformation. Prescribing the deformation only locally allows to deform the rest of the shape according to some prior, such as \emph{semantics} or \emph{deformation} rigidity. Different to previous efforts, our method is model-agnostic and can be applied to a pre-trained NN and update it in-place. Furthermore, we show how boundary sensitivity helps 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
- abstract@open-review(Poster): 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.

Revisiting Populations in multi-agent Communication

- Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou
- abstract@open-review(Poster): 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).

Sequential Gradient Coding For Straggler Mitigation

- Nikhil Krishnan Muralee Krishnan, MohammadReza Ebrahimi, Ashish J Khisti
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): This paper proposes a novel batch normalization strategy for test-time adaptation. Recent test-time adaptation methods heavily rely on the modified batch normalization, 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 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 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.

Disentanglement of Correlated Factors via Hausdorff Factorized Support

- Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, Diane Bouchacourt
- abstract@open-review(Poster): 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 representations 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 propose a relaxed disentanglement criterion – the Hausdorff Factorized Support (HFS) criterion – that encourages a 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.

Generating Sequences by Learning to Self-Correct

- Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, Yejin Choi
- abstract@open-review(Poster): 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.

Interneurons accelerate learning dynamics in recurrent neural networks for statistical adaptation

- David Lipshutz, Cengiz Pehlevan, Dmitri Chklovskii
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Latent Neural ODEs with Sparse Bayesian Multiple Shooting

- Valerii Iakovlev, Cagatay Yildiz, Markus Heinonen, Harri Lähdesmäki
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, Geunbae Lee, Mulong Luo, Peter Stone, Hsien-Hsin S. Lee, Benjamin Lee, G. Edward Suh, Wenjie Xiong, Yuandong Tian
- abstract@open-review(Poster): Security vulnerabilities in computer systems raise serious concerns as computers process an unprecedented amount of private and sensitive data today. Cache-timing attacks pose an important practical threat as they have been shown to be able to effectively breach many protection mechanisms in today's system. However, the current detection of cache timing attacks relies heavily on heuristics and expert knowledge, which can lead to brittleness and inability to adapt to new attacks. To mitigate these problems, we develop a two-player environment for cache-timing attacks and detection, and leverage the idea of population-based multi-agent reinforcement learning (MARL) to train both attackers and detectors. Our empirical results indicate that, without any manual input from security experts, the trained attacker is able to act more stealthily while the trained detector can generalize to \emph{unseen} attacks and is less exploitable to high-bandwidth attacks. Furthermore, in this environment, we found that agents equipped with a Transformer encoder substantially outperform agents with multi-layer perceptrons encoders, which has been commonly used in RL tasks, suggesting that Transformer may learn better representations in such real-world tasks.

PAC Reinforcement Learning for Predictive State Representations

- Wenhao Zhan, Masatoshi Uehara, Wen Sun, Jason D. Lee
- abstract@open-review(Poster): 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 can 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.

Decentralized Optimistic Hyperpolicy Mirror Descent: Provably No-Regret Learning in Markov Games

- Wenhao Zhan, Jason D. Lee, Zhuoran Yang
- abstract@open-review(Poster): 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.

Robust Scheduling with GFlowNets

- David W Zhang, Corrado Rainone, Markus Peschl, Roberto Bondesan
- abstract@open-review(Poster): 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.

Autoregressive Conditional Neural Processes

- Wessel Bruinsma, Stratis Markou, James Requeima, Andrew Y. K. Foong, Anna Vaughan, Tom Andersson, Anthony Buonomo, Scott Hosking, Richard E Turner
- abstract@open-review(Poster): 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.

kNN Prompting: Learning Beyond the Context with Nearest Neighbor Inference

- Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, Yongdong Zhang
- abstract@open-review(Poster): In-Context Learning, which formulates target tasks as prompt completion conditioned on in-context demonstrations, has become the prevailing and standard utilization of large language models. In this paper, we disclose an actual predicament for this typical usage that it can not scale up with training data due to context length restrictions. We then advocate a simple and effective solution, kNN Prompting, which not only outperforms In-Context Learning under few shot scenarios, but more importantly, can scale up with as many training data as are available. kNN Prompting queries LLM with training data for distributed representations and caches them locally as anchors. At inference time, it predicts by simply aggregating nearest neighbors. We conduct comprehensive experiments and ablations across different scales of LLMs to demonstrate its substantial improvements, as well as other appealing aspects such as robustness and explainability. The proposed approach successfully bridges data scaling into model scaling, and brings new potentials for the gradient-free paradigm of LLM deployment.

Obtaining More Generalizable Fair Classifiers on Imbalanced Datasets

- Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, James Zou
- abstract@open-review(Poster): Imposing fairness constraints during learning has been widely used to ensure algorithmic fairness. However, many datasets have an inherent imbalance in certain label classes (e.g. "healthy") and sensitive subgroups (e.g. "older patients"), which 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 constraint is far from being satisfied for new users. In this paper, we propose a theoretically principled, yet flexible approach that 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, our approach can be directly applied to achieve equalized opportunity (EqOpt); under certain conditions, it can also be applied to other fairness notions. We

demonstrate the power of our new approach by combining it with a popular fair classification algorithm, and the resulting algorithm achieves significantly better fairness generalization on several real-world datasets.

Understanding The Robustness of Self-supervised Learning Through Topic Modeling

- Zeping Luo, Shiyou Wu, Cindy Weng, Mo Zhou, Rong Ge
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, which 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, we also find 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).

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
- abstract@open-review(Poster): 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.

Strong inductive biases provably prevent harmless interpolation

- Michael Aerni, Marco Milanta, Konstantin Donhauser, Fanny Yang
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Recently, a lot of advances in self-supervised visual learning are brought about by contrastive learning that aligns positive pairs while pushing negative pairs apart. Surprisingly, a variety of new methods, such as BYOL, SimSiam, SwAV, DINO, shows that when equipped with some architectural asymmetric designs, aligning positive pairs alone is sufficient to attain good performance. However, it is still not fully clear how these seemingly different asymmetric designs can avoid feature collapse. Despite some understandings of some specific modules (like the predictor in BYOL), there is yet no unified theoretical understanding, particularly for those who also work without the predictor (like DINO). In this work, we propose a new understanding for non-contrastive learning, named the Rank Differential Mechanism (RDM). We show that these asymmetric designs all create a consistent difference in the dual-branch outputs as measured by their effective rank. 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.

Stay Moral and Explore: Learn to Behave Morally in Text-based Games

- Zijing Shi, Meng Fang, Yunqiu Xu, Ling Chen, Yali Du
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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).

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
- abstract@open-review(Poster): 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.

Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning

- Zhendong Wang, Jonathan J Hunt, Mingyuan Zhou
- abstract@open-review(Poster): 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.

Achieve 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
- abstract@open-review(Poster): Cooperative multi-agent multi-armed bandits (\CMAB) study how distributed agents cooperatively play the same multi-armed bandit game. Most existing \CMAB 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, abad" 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 \CMAB, 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 paper, 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Efficient Learning of Rationalizable Equilibria in General-Sum Games

- Yuanhao Wang, Dingwen Kong, Yu Bai, Chi Jin
- abstract@open-review(Poster): 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 profile 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.

Energy-Based Test Sample Adaptation for Domain Generalization

- Zehao Xiao, Xiantong Zhen, Shengcai Liao, Cees G. M. Snoek
- abstract@open-review(Poster): 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.

Bidirectional Language Models Are Also Few-shot Learners

- Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, Chris Callison-Burch
- abstract@open-review(Poster): 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.

EPISODE: Episodic Gradient Clipping with Periodic Resampled Corrections for Federated Learning with Heterogeneous Data

- Michael Crawshaw, Yajie Bao, Mingrui Liu
- abstract@open-review(Poster): 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.

A Theory of Dynamic Benchmarks

- Ali Shirali, Rediet Abebe, Moritz Hardt
- abstract@open-review(Poster): 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 setting 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
- abstract@open-review(Poster): 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 linear models and overparameterized neural networks – studied through the lens of neural tangent kernels (NTKs) – 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 What and Where - Unsupervised Disentangling Location and Identity Tracking

- Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thuemmel, Martin V. Butz
- abstract@open-review(Poster): 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 encodings, 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-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 an encoding that will facilitate effective planning and reasoning on conceptual levels.

BALTO: efficient 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
- abstract@open-review(Poster): 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.

Computing all Optimal Partial Transports

- Abhijeet Phatak, Sharath Raghvendra, CHITTARANJAN TRIPATHY, Kaiyi Zhang
- abstract@open-review(Poster): 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 mass distributions with $S \le U$. 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. OT-profile has been used for studying mathematical properties of optimal partial transports (see~\cite{figalli2010optimal}). In this paper, we consider the question of computing the OT-profile. 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 complexity\footnote{The combinatorial complexity of such a piecewise-linear function is simply the number of line segments it contains.} of this function, we present an exact algorithm to compute the OT-profile in $\tilde{O}(n^2K)$ time. Given $\delta > 0$, we also show that the algorithm by (Lahn et al., NeurIPS 2019) can be used to δ-approximate the OT-profile of μ and ν with another piecewise-linear function in $O(n^2/\delta + n/\delta^2)$ time. The complexity of this approximation is just $O(1/\delta)$. An OT-profile is arguably more valuable than the OT-cost itself and can be used within applications. For instance, under a reasonable assumption of outliers, we prove that the first derivative of the OT-profile sees a noticable rise before any of the mass from outliers is transported. By using the OT-profile, we get an improved prediction accuracy for an outlier detection experiment as well as estimation of class priors within PU-Learning experiments, both of which are conducted on real-datasets.

AE-FLOW: Autoencoders with Normalizing Flows for Medical Images Anomaly Detection

- Yuzhong Zhao, Qiaoqiao Ding, Xiaoqun Zhang
- abstract@open-review(Poster): 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.

A Self-Attention Ansatz for Ab-initio Quantum Chemistry

- Ingrid von Glehn, James S Spencer, David Pfau
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Pseudo-label Training and Model Inertia in Neural Machine Translation

- Benjamin Hsu, Anna Currey, Xing Niu, Maria Nadejde, Georgiana Dinu
- abstract@open-review(Poster): 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.

HyperDeepONet: learning operator with complex target function space using the limited resources via hypernetwork

- Jae Yong Lee, SungWoong CHO, Hyung Ju Hwang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 are unable to 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 `ignoreglobal' 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. By doing so, it 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 ECGAN achieves significantly better results than state-of-the-art methods.

CANIFE: Crafting Canaries for Empirical Privacy Measurement in Federated Learning

- Samuel Maddock, Alexandre Sablayrolles, Pierre Stock
- abstract@open-review(Poster): 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 not directly translate to 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 epsilon obtained with CANIFE is 2-7$\times$ lower than the theoretical bound.

A View From Somewhere: Human-Centric Face Representations

- Jerone Theodore Alexander Andrews, Przemyslaw Joniak, Alice Xiang
- abstract@open-review(Poster): Biases in human-centric computer vision models are often attributed to a lack of sufficient data diversity, with many demographics insufficiently represented. However, auditing datasets for diversity can be difficult, due to an absence of ground-truth labels of relevant features. 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. Moreover, categorical demographic labels do not necessarily capture all the relevant dimensions of human diversity that are important for developing fair and robust models. 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 novel dataset of 638,180 human judgments of face similarity (FAX). 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 importance they place on different attributes when judging similarity, underscoring the need for diverse annotator groups to avoid biases.

Identifiability Results for Multimodal Contrastive Learning

- Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, Julia E Vogt
- abstract@open-review(Poster): 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, Eric Xing, Yoon Kim
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 datasets, and outperforms the original dDGM model.

This Looks Like It Rather Than That: ProtoKNN For Similarity-Based Classifiers

- Yuki Ukai, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi
- abstract@open-review(Poster): 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.

Understanding weight-magnitude hyperparameters in training binary networks

- Joris Quist, Yunqiang Li, Jan van Gemert
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Contrastive Meta-Learning for Partially Observable Few-Shot Learning

- Adam Jelley, Amos Storkey, Antreas Antoniou, Sam Devlin
- abstract@open-review(Poster): 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 Dynamical Systems

- Suresh Bishnoi, Ravinder Bhattoo, Jayadeva Jayadeva, Sayan Ranu, N M Anoop Krishnan
- abstract@open-review(Poster): Neural networks with physics-based inductive biases such as Lagrangian neural networks (Lnn), and Hamiltonian neural networks (Hnn) 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 analyse 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. 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 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 $\sim$4 orders of magnitude for a pendulum system, and $\sim$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

- abstract@open-review(Poster): 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.

Correlative Information Maximization Based Biologically Plausible Neural Networks for Correlated Source Separation

- Bariscan Bozkurt, Ateş İsfendiyaroğlu, Cengiz Pehlevan, Alper Tunga Erdogan
- abstract@open-review(Poster): 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.

Leveraging Importance Weights in Subset Selection

- Gui Citovsky, Giulia DeSalvo, Sanjiv Kumar, Srikumar Ramalingam, Afshin Rostamizadeh, Yunjuan Wang
- abstract@open-review(Poster): 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 sampling 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.

Copy is All You Need

- Tian Lan, Deng Cai, Yan Wang, Heyan Huang, Xian-Ling Mao
- abstract@open-review(Poster): The dominant text generation models compose output by selecting words in 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 existing articles in 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 by coping from the original training data (0.758 vs. 0.691 MAUVE). We also show that our approach attains additional performance gains by simply scaling up to larger text collections without extra training. Furthermore, our approach allows for effective domain adaptation by simply switching to any domain-specific text collection, again without further training. Finally, we observe that our approach achieves better inference efficiency than standard token-level autoregressive models thanks to the reduction of decoding steps.

Why adversarial training can hurt robust accuracy

- Jacob Clarysse, Julia Hörrmann, Fanny Yang
- abstract@open-review(Poster): 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.

Representational Dissimilarity Metric Spaces for Stochastic Neural Networks

- Lyndon Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, Alex H Williams
- abstract@open-review(Poster): 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.

Sequential Learning of Neural Networks for Prequential MDL

- Jorg Bornschein, Yazhe Li, Marcus Hutter
- abstract@open-review(Poster): 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 previously 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.

Learning topology-preserving data representations

- Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Serguei Barannikov, Evgeny Burnaev
- abstract@open-review(Poster): 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 representation 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.

The Curious Case of Benign Memorization

- Sotiris Anagnostidis, Gregor Bachmann, Lorenzo Noci, Thomas Hofmann
- abstract@open-review(Poster): 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.

Unbiased Supervised Contrastive Learning

- Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, Pietro Gori
- abstract@open-review(Poster): 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".

Compositional Prompt Tuning with Motion Cues for Open-vocabulary Video Relation Detection

- Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, Qianru Sun
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 we propose to approximate the hypervolume function with a deep neural network, 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 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 benchmark test cases. The results show that our methods are promising for such multi-objective optimization tasks.

DiffusER: Diffusion via Edit-based Reconstruction

- Machel Reid, Vincent Josua Hellendoorn, Graham Neubig
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning to Solve Constraint Satisfaction Problems with Recurrent Transformers

- Zhun Yang, Adam Ishay, Joohyung Lee
- abstract@open-review(Poster): Constraint satisfaction problems (CSPs) are about finding values of variables that satisfy the given constraints. We show that the Transformer model extended with recurrence is a viable approach to learning to solve CSPs in an end-to-end manner, having clear advantages over the state-of-the-art methods such as Graph Neural Networks, SATNet, and some neuro-symbolic models. With the ability of Transformers 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.

Improving the imputation of missing data with Markov Blanket discovery

- Yang Liu, Anthony Constantinou
- abstract@open-review(Poster): 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 incorporate 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
- abstract@open-review(Poster): 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.

Fundamental Limits in Formal Verification of Message-Passing Neural Networks

- Marco Sälzer, Martin Lange
- abstract@open-review(Poster): 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 expressive 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.

Short-Term Memory Convolutions

- Grzegorz Stefański, Krzysztof Arendt, Paweł Daniluk, Bartłomiej Jasik, Artur Szumaczuk
- abstract@open-review(Poster): 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 for audio domain as the human audio perception 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 a method for minimization of 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 an audio separation task. We achieved a 5-fold reduction in inference time and a 2-fold reduction in latency without affecting the output quality.

LexMAE: Lexicon-Bottlenecked Pretraining for Large-Scale Retrieval

- Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang
- abstract@open-review(Poster): 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

- abstract@open-review(Poster): 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 predict the solution distributions, then we search for the best feasible solution within a properly defined ball around the predicted solution. We show that our framework is both theoretically and computationally superior to fixing strategies. We conduct extensive experiments on four public datasets and numerical results demonstrate that our proposed framework achieves 51% and 14% better primal gaps than state-of-the-art general-purpose optimization solvers SCIP and Gurobi, respectively.

On Explaining Neural Network Robustness with Activation Path

- Ziping Jiang
- abstract@open-review(Poster): 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 classifier 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.

Structure by Architecture: Structured Representations without Regularization

- Felix Leeb, Giulia Lanzillotta, Yashas Annadani, Michel Besserve, Stefan Bauer, Bernhard Schölkopf
- abstract@open-review(Poster): 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 distribution 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
- abstract@open-review(Poster): 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 leads to better-behaved optimization. We also propose a solution to the problem of $\textit{ensemble detection}$, whereby different groups of neurons, i.e., ensembles, can be modulated by different latent manifolds. This is achieved through a soft clustering of neurons during training, thus allowing 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, recovering distinct ensembles, inferring toroidal latents and predicting neural tuning curves all in a single integrated modeling framework.

Learning Fast and Slow for Time Series Forecasting

- Quang Pham, Chenghao Liu, Doyen Sahoo, Steven Hoi
- abstract@open-review(Poster): 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 will be made available.

Guess the Instruction! Making Language Models Stronger Zero-Shot Learners

- Seonghyeon Ye, Doyoung Kim, Joel Jang, Joongbo Shin, Minjoon Seo
- abstract@open-review(Poster): 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 3B-sized Flipped outperforms 4 times larger zero-shot T0-11B (Sanh et al, 2021) and even a 60 times larger 3-shot GPT-3 (175B) (Brown et al, 2020) on average by 1.8% and 3.1%, respectively. Flipped gives particularly large improvements on 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.

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
- abstract@open-review(Poster): 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 intro- duce 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.

DECAP: Decoding CLIP Latents for Zero-shot Captioning

- Wei Li, Linchao Zhu, Longyin Wen, Yi Yang
- abstract@open-review(Poster): 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 interests. 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 computationally-efficient: 1) it only requires the 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 only the text 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 em-bedding and the visual embedding are correlated, the 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 by a large margin on the typical image captioning benchmarks, i.e., MSCOCO and Flickr30k. On Flickr30k, the zero-shot result is even competitive with fully supervised methods. We apply DeCap to video captioning and achieve state-of-the-art zero-shot performance on MSR-VTT and ActivityNet-Captions.

That Label's got Style: Handling Label Style Bias for Uncertain Image Segmentation

- Kilian Zepf, Eike Petersen, Jes Frellsen, Aasa Feragen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Neural networks can be drastically shrinked 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, learning a global compression strategy that, however, can be specific to each layer. Our method optimizes both how many parameters (compression rate) and which parameters (scoring connections) to prune. It 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 as well as its adversarially robustness for a reduction by 99% of the network's original size. Moreover, we observe a crucial influence of non-uniform compression across layers.

PASHA: Efficient HPO and NAS with Progressive Resource Allocation

- Ondrej Bohdal, Lukas Balles, Martin Wistuba, Beyza Ermis, Cedric Archambeau, Giovanni Zappella
- abstract@open-review(Poster): 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.

StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random

- Haoxuan Li, Chunyuan Zheng, Peng Wu
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 +54% 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.

Theoretical Characterization of the Generalization Performance of Overfitted Meta-Learning

- Peizhong Ju, Yingbin Liang, Ness Shroff
- abstract@open-review(Poster): 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 large. Under this circumstance, we show that the overfitted min ℓ_2-norm solution can achieve an even lower generalization error than the underparameterized solution.

Robust Explanation Constraints for Neural Networks

- Matthew Robert Wicker, Juyeon Heo, Luca Costabello, Adrian Weller
- abstract@open-review(Poster): 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.

Offline Reinforcement Learning via High-Fidelity Generative Behavior Modeling

- Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, Jun Zhu
- abstract@open-review(Poster): 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.

CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers

- Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, Jie Tang
- abstract@open-review(Poster): Large-scale pretrained transformers have reached a milestone in text (GPT-3) and text-to-image (DALL-E and CogView) generation. However, its application to video generation still has several challenges: unaffordable huge computation cost and scarcity and weak relevance of the text-video datasets. 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. The anonymous web demo is available at https://cogvideo.pka.moe.

Revisit Finetuning strategy for Few-Shot Learning to Strengthen the Equivariance of Emdeddings

- Heng Wang, Tan Yue, Xiang Ye, Zihang He, Bohan Li, Yong Li
- abstract@open-review(Poster): 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, imposing equivariance on the whole model. 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.

Optimizing Spca-based Continual Learning: A Theoretical Approach

- Chunchun Yang, Malik Tiomoko, Zengfu Wang
- abstract@open-review(Poster): 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

- abstract@open-review(Poster): 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 will be made publicly available.

Sampling-free Inference for Ab-Initio Potential Energy Surface Networks

- Nicholas Gao, Stephan Günnemann
- abstract@open-review(Poster): 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%.

A New Hierarchy of Expressivity for Graph Neural Networks

- Qing Wang, Dillon Ze Chen, Asiri Wijesinghe, Shouheng Li, Muhammad Farhan
- abstract@open-review(Poster): 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 testing. 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 that go beyond 1-WL. Thus, we propose a novel hierarchy of graph isomorphism tests, namely \emph{Neighbourhood WL} ($\mathscr{N}$-WL) algorithms, which enables to better measure the expressive power of GNNs. We further introduce \emph{Graph Neighbourhood Neural Network} (G3N) by building upon the $\mathscr{N}$-WL algorithms, 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
- abstract@open-review(Poster): 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.

DAVA: Disentangling Adversarial Variational Autoencoder

- Benjamin Estermann, Roger Wattenhofer
- abstract@open-review(Poster): The use of well-disentangled representations poses many advantages for downstream tasks, e.g. increasing sample efficiency, or enabling interpretability. Their quality is, however, determined to a large extent by the choice of dataset-specific hyperparameters, most notably the regularization strength. To address the issue, we introduce DAVA, a novel training procedure for variational auto-encoders that alleviates the issue of hyperparameter selection at the cost of a comparatively small overhead. We compare DAVA against models with optimal choice of hyperparameters. Without any hyperparameter tuning, DAVA is competitive across a diverse range of commonly used datasets. Further, even under an adequate set of hyperparameters, the success of the disentanglement process remains heavily influenced by randomness in network initialization. We therefore present the new unsupervised PIPE disentanglement metric, capable of evaluating representation quality. We demonstrate the PIPE metrics ability to positively predict performance of downstream models in abstract reasoning. We also exhaustively examine correlations with existing supervised and unsupervised metrics.

TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations

- Haoxuan Li, Yan Lyu, Chunyuan Zheng, Peng Wu
- abstract@open-review(Poster): 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 often occur in practice. In this paper, we propose a principled approach that can effectively reduce the bias and variance simultaneously for existing DR estimators when the error-imputation model is misspecified. In addition, we further propose a novel semi-parametric collaborative counterfactual 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.

Domain Generalisation via Domain Adaptation: An Adversarial Fourier Amplitude Approach

- Minyoung Kim, Da Li, Timothy Hospedales
- abstract@open-review(Poster): 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.

Consolidator: Mergable Adapter with Group Connections for Vision Transformer

- Tianxiang Hao, Hui Chen, Yuchen Guo, Guiguang Ding
- abstract@open-review(Poster): 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 of accommodating numerous parameters. As a result, new challenges for adapting a well-trained transformer to downstream tasks arise. On resource-limited devices, classic fine-tuning, which tunes and stores a full copy of parameters in the pretrained model for every downstream task, is usually impracticable for the shortage of storage space. However, few works have focused on how to efficiently and effectively transfer the 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 vision transformers. Our consolidator modifies the pretrained 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 transformer blocks to construct tunable parts in a consolidator. To further enhance the model capacity to transfer knowledge under constrained storage budget and keep inference efficient, we consolidate the parameters in two stages: 1. between adaptation and storage 2. between loading and inference. On a series of downstream visual tasks, our consolidator can reach better performance than full fine-tuning with less than 0.5\% parameters stored per task, and outperform state-of-the-art parameter-efficient tuning methods.

Statistical Theory of Differentially Private Marginal-based Data Synthesis Algorithms

- Ximing Li, Chendi Wang, Guang Cheng
- abstract@open-review(Poster): 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.

Anti-Symmetric DGN: a stable architecture for Deep Graph Networks

- Alessio Gravina, Davide Bacciu, Claudio Gallicchio
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 the 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.

Online Low Rank Matrix Completion

- Soumyabrata Pal, Prateek Jain
- abstract@open-review(Poster): 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.

Explaining RL Decisions with Trajectories

- Shripad Vilasrao Deshmukh, Arpan Dasgupta, Chirag Agarwal, Nan Jiang, Balaji Krishnamurthy, Georgios Theocharous, Jayakumar Subramanian
- abstract@open-review(Poster): 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.

FastFill: Efficient Compatible Model Update

- Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, Hadi Pouransari

- abstract@open-review(Poster): 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 \emph{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 \method: 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 \emph{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
- abstract@open-review(Poster): Existing Graph Neural Networks (GNNs) compute the message exchange between nodes by either aggregating uniformly (convolving) the features of all the neighbor- ing nodes, or by applying a non-uniform score (attending) to the features. Recent works have shown the strengths and weaknesses of the resulting GNN architectures, 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.

Scaffolding a Student to Instill Knowledge

- Anil Kag, Durmus Alp Emre Acar, Aditya Gangrade, Venkatesh Saligrama
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 important hyperparameter - the proximity of the learned policy to the original policy. We propose an algorithm that allows the user to tune this hyperparameter at runtime, 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.

SLTUNET: A Simple Unified Model for Sign Language Translation

- Biao Zhang, Mathias Müller, Rico Sennrich
- abstract@open-review(Poster): 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 will be released.

Understanding the Generalization of Adam in Learning Neural Networks with Proper Regularization

- Difan Zou, Yuan Cao, Yuanzhi Li, Quanquan Gu
- abstract@open-review(Poster): 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.

A law of adversarial risk, interpolation, and label noise

- Daniel Paleka, Amartya Sanyal
- abstract@open-review(Poster): 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.

Learning ReLU networks to high uniform accuracy is intractable

- Julius Berner, Philipp Grohs, Felix Voigtlaender

- abstract@open-review(Poster): 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 depth and the input dimension of the network architecture.

Active Learning for Object Detection with Evidential Deep Learning and Hierarchical Uncertainty Aggregation

- Younghyun Park, Soyeong Kim, Wonjeong Choi, Dong-Jun Han, Jaekyun Moon
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 another notion of sharpness, which is most likely to be the preferred notion for practical performance. The key mechanism behind this intriguing phenomenon is the implicit alignment between the gradient and the top eigenvector of Hessian when running SAM.

The Implicit Bias of Minima Stability in Multivariate Shallow ReLU Networks

- Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, Daniel Soudry
- abstract@open-review(Poster): 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) f, whose second derivative has a bounded weighted L^1 norm. The bound on the norm 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, Δf. 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. In contrast, we show that the same function can be realized as a stable two hidden-layer ReLU network. Finally, we prove that if a function is sufficiently smooth (Sobolev) then it can be approximated arbitrarily well using single hidden-layer ReLU networks that correspond to stable solutions of gradient descent.

MAST: Masked Augmentation Subspace Training for Generalizable Self-Supervised Priors

- Chen Huang, Hanlin Goh, Jiatao Gu, Joshua M. Susskind
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Wasserstein Auto-encoded MDPs: Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees

- Florent Delgrange, Ann Nowe, Guillermo Perez
- abstract@open-review(Poster): 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 (mode 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.

Global Explainability of GNNs via Logic Combination of Learned Concepts

- Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Lio, Andrea Passerini
- abstract@open-review(Poster): 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 debugging. 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.

Gradient Gating for Deep Multi-Rate Learning on Graphs

- T. Konstantin Rusch, Benjamin Paul Chamberlain, Michael W. Mahoney, Michael M. Bronstein, Siddhartha Mishra
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Open-ended learning methods that automatically generate a curriculum of increasingly challenging tasks serve as a promising avenue toward generally capable reinforcement learning (RL) 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 then introduce Multi-Agent Environment-Space Response Oracles (MAESTRO), the first multi-agent UED approach for two-player zero-sum settings. MAESTRO efficiently produces adversarial, joint curricula over both environment parameters and co-player policies and attains minimax-regret guarantees at Nash equilibrium. Our experiments show that MAESTRO outperforms a number of strong baselines on competitive two-player environments, spanning discrete and continuous control.

Almost Linear Constant-Factor Sketching for ℓ_1 and Logistic Regression

- Alexander Munteanu, Simon Omlor, David Woodruff
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Leveraging Unlabeled Data to Track Memorization

- Mahsa Forouzesh, Hanie Sedghi, Patrick Thiran
- abstract@open-review(Poster): 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.

Policy-Based Self-Competition for Planning Problems

- Jonathan Pirnay, Quirin Göttl, Jakob Burger, Dominik Gerhard Grimm

- abstract@open-review(Poster): AlphaZero-type algorithms may stop improving on single-player tasks in case the value network guiding the tree search is unable to approximate the exact 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 Scheduling Problem. With only half of the simulation budget for search, GAZ PTP consistently outperforms all selected single-player variants of GAZ.

Efficient 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
- abstract@open-review(Poster): Out-of-Distribution (OOD) detection is essential for safety-critical applications of deep neural networks. OOD detection is challenging since DNN models tend to 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. Unlike existing OOD methods refining the confidence estimation procedure from output logits with handpicked hyperparameters, we propose a new store-then-compare paradigm. In more detail, penultimate layer outputs on the training set are considered as the representation 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 an energy function defined in Modern Hopfield Network for the discrepancy score calculation, we derive a simplified version SHE with theoretical analysis. In SHE, we only utilize 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 hyperparameter-free 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.

Pareto-Efficient Decision Agents for Offline Multi-Objective Reinforcement Learning

- Baiting Zhu, Meihua Dang, Aditya Grover
- abstract@open-review(Poster): 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 demonstrations 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.

NAGphormer: A Tokenized Graph Transformer for Node Classification in Large Graphs

- Jinsong Chen, Kaiyuan Gao, Gaichao Li, Kun He
- abstract@open-review(Poster): 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 quadratical 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.

Bayesian Oracle for bounding information gain in neural encoding models

- Konstantin-Klemens Lurz, Mohammad Bashiri, Edgar Y. Walker, Fabian H. Sinz
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Learning Vortex Dynamics for Fluid Inference and Prediction

- Yitong Deng, Hong-Xing Yu, Jiajun Wu, Bo Zhu
- abstract@open-review(Poster): We propose a novel machine learning method based on differentiable vortex particles to infer and predict fluid dynamics from a single video. The key design of our system is a particle-based latent space to encapsulate the hidden, Lagrangian vortical evolution underpinning the observable, Eulerian flow phenomena. We devise a novel differentiable vortex particle system in conjunction with their learnable, vortex-to-velocity dynamics mapping to effectively capture and represent the complex flow features in a reduced space. We further design an end-to-end training pipeline to directly learn and synthesize simulators from data, that can reliably deliver future video rollouts based on limited observation. 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, to be used for motion analysis; secondly, it also supports future prediction, constructing the input video's sequel along with its future dynamics evolution. We demonstrate our method's efficacy by comparing quantitatively and qualitatively with a range of existing methods on both synthetic and real-world videos, displaying improved data correspondence, 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
- abstract@open-review(Poster): 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.

Quality-Similar Diversity via Population Based Reinforcement Learning

- Shuang Wu, Jian Yao, Haobo Fu, Ye Tian, Chao Qian, Yaodong Yang, QIANG FU, Yang Wei
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Tensor-Based Sketching Method for the Low-Rank Approximation of Data Streams.

- Cuiyu Liu, Xiao Chuanfu, Mingshuo Ding, Chao Yang
- abstract@open-review(Poster): 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.

Language Models are Realistic Tabular Data Generators

- Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, Gjergji Kasneci
- abstract@open-review(Poster): 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 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 many real-world data sets with heterogeneous feature types coming in various sizes.

Data augmentation alone can improve adversarial training

- Lin Li, Michael W. Spratling
- abstract@open-review(Poster): 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.

CUTS: Neural Causal Discovery from Unstructured Time-Series Data

- Cheng Yuxiao, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, Qionghai Dai
- abstract@open-review(Poster): 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 unstructured 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.

Quantized Compressed Sensing with Score-Based Generative Models

- Xiangming Meng, Yoshiyuki Kabashima
- abstract@open-review(Poster): We consider the general problem of recovering a high-dimensional signal from noisy quantized measurements. Quantization, especially coarse quantization such as one-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 likelihood score called noise perturbed likelihood score is introduced and combined with the prior score of SGM. The proposed QCS-SGM applies to 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. Our code will be open-sourced after acceptance.

Valid P-Value for Deep Learning-driven Salient Region

- Miwa Daiki, Vo Nguyen Le Duy, Ichiro Takeuchi
- abstract@open-review(Poster): 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.

Complexity-Based Prompting for Multi-step Reasoning

- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, Tushar Khot
- abstract@open-review(Poster): 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.

Unsupervised 3d object learning through neuron activity aware plasticity

- Beomseok Kang, Biswadeep Chakraborty, Saibal Mukhopadhyay
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Incremental Learning of Structured Memory via Closed-Loop Transcription

- Shengbang Tong, Xili Dai, Ziyang Wu, Mingyang Li, Brent Yi, Yi Ma
- abstract@open-review(Poster): 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.

When Data Geometry Meets Deep Function: Generalizing Offline Reinforcement Learning

- Jianxiong Li, Xianyuan Zhan, Haoran Xu, Xiangyu Zhu, Jingjing Liu, Ya-Qin Zhang
- abstract@open-review(Poster): 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.

Budgeted Training for Vision Transformer

- zhuofan xia, Xuran Pan, Xuan Jin, Yuan He, Hui Xue', Shiji Song, Gao Huang
- abstract@open-review(Poster): 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 any training budget, while achieving competitive model performances. Specifically, based on the observation that Transformer exhibits different levels of model redundancies at different stages of training, we propose to dynamically control the activation rate of model parameters 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.

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
- abstract@open-review(Poster): 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.

What Do Self-Supervised Vision Transformers Learn?

- Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, Sangdoo Yun
- abstract@open-review(Poster): We present comparative studies on how and why contrastive learning (CL) and masked image modeling (MIM) differ in their representations and performance on downstream tasks. In particular, self-supervised Vision Transformers (ViTs) have the following properties: (1) CL trains self-attentions to capture longer-range global patterns, such as the shape of an object, compared with MIM. This property of CL helps ViTs linearly separate images in their representation spaces. However, it also makes the self-attentions collapse into homogeneity for all heads, depths, and query tokens. Such homogeneity of self-attention reduces representations' diversity, resulting in worse scalability and dense prediction performance; (2) CL reduces the high-frequency signals of the representations, but MIM amplifies them. Since the low-frequency information stands for the shapes and the high frequencies represent the textures, CL is more shape-oriented, whereas MIM is more texture-oriented; (3) CL plays a crucial role in the later layers of ViT architecture, while MIM mainly focuses on the early layers. Upon these analyses, we find that CL and MIM can complement each other and observe that the simplest harmonization can enjoy the advantages of both methods.

Scaling Laws in Mean-Field Games

- Pengdeng Li, Xinrun Wang, Shuxin Li, Hau Chan, Bo An
- abstract@open-review(Poster): In this work, we attempt to bridge the two largely independently evolving fields of finite-agent and infinite-agent games, by studying the scaling laws in mean-field games. The key is to obtain the optimal policies of a set of finite-agent games with different numbers of agents (population size) and then investigate how the policies evolve with the population size. 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 extensive analysis of the scaling laws of the generated policies further deepens our understanding of the two fields of finite-agent and infinite-agent games. To our best knowledge, this work presents the first attempt to bridge the two research fields.

On The Relative Error of Random Fourier Features for Preserving Kernel Distance

- Kuan Cheng, Shaofeng H.-C. Jiang, Luojian Wei, Zhide Wei
- abstract@open-review(Poster): 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.

NewModel: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing

- Pengcheng He, Jianfeng Gao, Weizhu Chen
- abstract@open-review(Poster): 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.

Squeeze Training for Adversarial Robustness

- Qizhang Li, Yiwen Guo, Wangmeng Zuo, Hao Chen
- abstract@open-review(Poster): The vulnerability of deep neural networks (DNNs) to adversarial examples has attracted great attention in the machine learning community. The problem is related to local non-smoothness and steepness 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 will be made publicly available.

Pushing the Accuracy-Fairness Tradeoff Frontier with Introspective Self-play

- Jeremiah Zhe Liu, Krishnamurthy Dj Dvijotham, Jihyeon Lee, Quan Yuan, Balaji Lakshminarayanan, Deepak Ramachandran
- abstract@open-review(Poster): Improving the accuracy-fairness frontier of deep neural network (DNN) models is an important problem. 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.

Max-Margin Works while Large Margin Fails: Generalization without Uniform Convergence

- Margalit Glasgow, Colin Wei, Mary Wootters, Tengyu Ma
- abstract@open-review(Poster): 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.

Asymptotic Instance-Optimal Algorithms for Interactive Decision Making

- Kefan Dong, Tengyu Ma
- abstract@open-review(Poster): 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.

Near-Optimal Deployment Efficiency in Reward-Free Reinforcement Learning with Linear Function Approximation

- Dan Qiao, Yu-Xiang Wang
- abstract@open-review(Poster): 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.

An Equal-Size Hard EM Algorithm for Diverse Dialogue Generation

- Yuqiao Wen, Yongchang Hao, Yanshuai Cao, Lili Mou
- abstract@open-review(Poster): Open-domain dialogue systems aim to interact with humans through natural language texts in an open-ended fashion. However, the widely successful neural networks may not work well for dialogue systems, as they tend to generate generic responses. 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.

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
- abstract@open-review(Poster): 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.

Long-Tailed Partial Label Learning via Dynamic Rebalancing

- Feng Hong, Jiangchao Yao, Zhihan Zhou, Yanfeng Wang, Ya Zhang
- abstract@open-review(Poster): 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. Our code will be publicly available.

Task Ambiguity in Humans and Language Models

- Alex Tamkin, Kunal Handa, Avash Shrestha, Noah Goodman
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

More Centralized Training, Still Decentralized Execution: Multi-Agent Conditional Policy Factorization

- Jiangxing Wang, Deheng Ye, Zongqing Lu
- abstract@open-review(Poster): 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.

Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks

- Bowen Jin, Yu Zhang, Yu Meng, Jiawei Han
- abstract@open-review(Poster): 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. Code can be found at https://anonymous.4open.science/r/Edgeformer-release-F422.

Any-scale Balanced Samplers for Discrete Space

- Haoran Sun, Bo Dai, Charles Sutton, Dale Schuurmans, Hanjun Dai
- abstract@open-review(Poster): 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.

Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design

- Keir Adams, Connor W. Coley
- abstract@open-review(Poster): Shape-based virtual screening is widely employed in ligand-based drug design to search chemical libraries for molecules with similar 3D shapes yet novel 2D chemical structures compared to known ligands. 3D deep generative models have the potential to automate this exploration of shape-

conditioned 3D chemical space; however, no existing models can reliably generate valid drug-like molecules in conformations that adopt a specific shape such as a known binding pose. 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 conformational structure 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.

Imbalanced Semi-supervised Learning with Bias Adaptive Classifier

- Renzhen Wang, Xixi Jia, Quanziang Wang, Yichen Wu, Deyu Meng
- abstract@open-review(Poster): 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.

On Compositional Uncertainty Quantification for Seq2seq Graph Parsing

- Zi Lin, Du Phan, Panupong Pasupat, Jeremiah Zhe Liu, Jingbo Shang
- abstract@open-review(Poster): 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.

Free Lunch for Domain Adversarial Training: Environment Label Smoothing

- YiFan Zhang, xue wang, Jian Liang, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Understanding Embodied Reference with Touch-Line Transformer

- Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, Yixin Zhu
- abstract@open-review(Poster): We study embodied reference understanding: locating referents using embodied gestural cues and language references. A popular misconception is that referents lie on the elbow-wrist line. However, as shown by human studies, the virtual touch line much more accurately indicates the direction of referents. This more accurate indicator of referent directions is missing from human pose representations in existing computational models. Consequently, existing computational models cannot effectively incorporate gestural information when locating referents. We help computational models utilize this critical gestural information by devising the touch-line transformer. Our touch-line transformer takes tokenized visual and textual features as inputs and simultaneously predicts the referent's bounding box and a touch-line vector. At the same time, to facilitate the use of the touchline prior, we apply a geometric consistency loss that encourages the co-linearity between referents and touch lines. Incorporating gestural information improves model performances significantly. Experiments on the YouRefIt dataset show our method achieves a +25.0% accuracy improvement under the 0.75 IoU criterion, closing 63.6% of the gap between model and human performances. Furthermore, we computationally verify prior human studies by showing that computational models more accurately locate referents when using the virtual touch line than when using the elbow-wrist line. Our codes and models will be publicly available.

Calibration Matters: Tackling Maximization Bias in Large-scale Advertising Recommendation Systems

- Yewen Fan, Nian Si, Kun Zhang
- abstract@open-review(Poster): 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.

Memorization-Dilation: Modeling Neural Collapse Under Noise

- Duc Anh Nguyen, Ron Levie, Julian Lienen, Eyke Hüllermeier, Gitta Kutyniok
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning to Extrapolate: A Transductive Approach

- Aviv Netanyahu, Abhishek Gupta, Max Simchowitz, Kaiqing Zhang, Pulkit Agrawal
- abstract@open-review(Poster): Machine learning systems, especially overparameterized deep neural networks, can generalize to novel testing instances drawn from the same distribution as the training data. However, they fare poorly when evaluated on out-of-support testing points. In this work, we tackle the problem of developing machine learning systems that retain the power of overparametrized function approximators, while enabling extrapolation to out-of-support testing 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. We instantiate a simple, practical algorithm applicable to various supervised learning problems and imitation learning tasks.

Label-free Concept Bottleneck Models

- Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, Tsui-Wei Weng
- abstract@open-review(Poster): Concept bottleneck model (CBM) are a popular way of creating more interpretable neural network by having hidden layer neurons correspond to human-understandable concepts. However, existing CBMs and their variants have two crucial limitations: first, the 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 adoption in practical real world applications. Motivated by these challenges, we propose \textit{Label-free} CBM which is a 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: \textit{scalable} - we present the first CBM scaled to ImageNet, \textit{efficient} - creating a CBM takes only a few hours even for very large datasets, and \textit{automated} - training it for a new dataset requires minimal human effort.

Multi-level Protein Structure Pre-training via Prompt Learning

- Zeyuan Wang, Qiang Zhang, Shuang-Wei HU, Haoran Yu, Xurui Jin, Zhichen Gong, Huajun Chen
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 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 disconvergence. 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://anonymous.4open.science/r/GLM-130B/.

Causal Estimation for Text Data with (Apparent) Overlap Violations

- Lin Gui, Victor Veitch
- abstract@open-review(Poster): 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.

MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations

- Nicklas Hansen, Yixin Lin, Hao Su, Xiaolong Wang, Vikash Kumar, Aravind Rajeswaran
- abstract@open-review(Poster): 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 260%-350% more successful in completing sparse reward tasks compared to prior approaches in the low data regime (100K interaction steps, 5 demonstrations). Webpage: https://modemrl.github.io/

PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm

- Toygun Basaklar, Suat Gumussoy, Umit Ogras
- abstract@open-review(Poster): 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.

Understanding the Role of Nonlinearity in Training Dynamics of Contrastive Learning

- Yuandong Tian
- abstract@open-review(Poster): 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.

M-L2O: Towards Generalizable Learning-to-Optimize by Test-Time Fast Self-Adaptation

- Junjie Yang, Xuxi Chen, Tianlong Chen, Zhangyang Wang, Yingbin Liang
- abstract@open-review(Poster): 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. All codes will be shared upon acceptance.

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
- abstract@open-review(Poster): Vision transformers (ViTs) have quickly superseded convolutional networks (ConvNets) as the current state-of-the-art (SOTA) models for medical image segmentation. Hierarchical transformers (e.g., Swin Transformers) reintroduced several ConvNet priors and further enhanced the practical viability of adapting volumetric 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 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).

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Explaining Temporal Graph Models through an Explorer-Navigator Framework

- Wenwen Xia, Mincai Lai, Caihua Shan, Yao Zhang, Xinnan Dai, Xiang Li, Dongsheng Li
- abstract@open-review(Poster): 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 tailored 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.

Soft Neighbors are Positive Supporters in Contrastive Visual Representation Learning

- Chongjian GE, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, Ping Luo
- abstract@open-review(Poster): 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 encoded 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. More materials can be found in our project page: anonymous-iclr23snclr.github.io.

Offline RL for Natural Language Generation with Implicit Language Q Learning

- Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, Sergey Levine
- abstract@open-review(Poster): 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.

CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos

- Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, Taylor Berg-Kirkpatrick
- abstract@open-review(Poster): 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 query 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
- abstract@open-review(Poster): 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.

An Adaptive Policy to Employ Sharpness-Aware Minimization

- Weisen Jiang, Hansi Yang, Yu Zhang, James Kwok
- abstract@open-review(Poster): 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 proposed method.

Fairness and Accuracy under Domain Generalization

- Thai-Hoang Pham, Xueru Zhang, Ping Zhang
- abstract@open-review(Poster): 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.

Language Models Can Teach Themselves to Program Better

- Patrick Haluptzok, Matthew Bowers, Adam Tauman Kalai
- abstract@open-review(Poster): 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.

Latent Bottlenecked Attentive Neural Processes

- Leo Feng, Hossein Hajimirsadeghi, Yoshua Bengio, Mohamed Osama Ahmed
- abstract@open-review(Poster): 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.

Embed to Control Partially Observed Systems: Representation Learning with Provable Sample Efficiency

- Lingxiao Wang, Qi Cai, Zhuoran Yang, Zhaoran Wang
- abstract@open-review(Poster): 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 Embed to Control (ETC), which learns the representation at two levels while optimizing the policy.~(i) For each step, ETC learns to represent the state with a low-dimensional feature, which factorizes the transition kernel. (ii) Across multiple steps, ETC 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, ETC 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, ETC 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
- abstract@open-review(Poster): 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 and confirm 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 probabilities 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.

Learning Kernelized Contextual Bandits in a Distributed and Asynchronous Environment

- Chuanhao Li, Huazheng Wang, Mengdi Wang, Hongning Wang
- abstract@open-review(Poster): 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 quality and communication efficiency. Rigorous theoretical analysis about the regret and communication cost is provided; and extensive empirical evaluations demonstrate the effectiveness of our solution.

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
- abstract@open-review(Poster): Inductive 1-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.

LipsFormer: Introducing Lipschitz Continuity to Vision Transformers

- Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, Lei Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Large language models (LLMs) can perform complex reasoning by generating intermediate reasoning steps. Providing such a series of steps for prompting demonstrations is called chain-of-thought (CoT) prompting. CoT prompting has two major paradigms. One leverages a simple prompt like ``Let's think step by step'' to facilitate step-by-step thinking before answering each question. The other uses a few step-by-step demonstrations, each composed of a question and a reasoning chain that leads to an answer. In practice, the second paradigm has achieved stronger performance than the first paradigm. However, this superior performance hinges on the hand-crafting of multiple effective task-specific demonstrations. We show that this limitation may be addressed by leveraging pre-existing abilities of LLMs to generate reasoning chains for demonstrations. A key challenge is that these generated chains often come with mistakes. To mitigate the effect of such mistakes, we investigate various principles for automatically constructing demonstrations and find that diversity matters. Inspired by these findings, we propose an automatic CoT prompting method called Auto-CoT. Auto-CoT samples questions with diversity and generates reasoning chains to construct demonstrations. On ten public benchmark reasoning tasks, Auto-CoT performs competitively compared to Manual-CoT that requires manual designs.

An efficient encoder-decoder architecture with top-down attention for speech separation

- Kai Li, Runxuan Yang, Xiaolin Hu
- abstract@open-review(Poster): 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 Sep- former. 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.

Machine Unlearning of Federated Clusters

- Chao Pan, Jin Sima, Saurav Prakash, Vishal Rana, Olgica Milenkovic
- abstract@open-review(Poster): Federated clustering 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 federated clustering methods has become of significant importance. This work proposes the first known unlearning mechanism for federated clustering with privacy criteria that support simple, provable, and efficient data removal at the client and server level. The gist of our approach is to combine special initialization procedures with quantization methods that allow for secure aggregation of estimated local cluster counts at the server unit. As part of our platform, we introduce secure compressed multiset aggregation (SCMA), which is of independent interest for secure sparse model aggregation. In order to simultaneously facilitate low communication complexity and secret sharing protocols, we integrate Reed-Solomon encoding with special evaluation points into the new SCMA pipeline and derive bounds on the time and communication complexity of different components of the scheme. Compared to completely retraining K-means++ locally and globally for each removal request, we obtain an average speed-up of roughly 84x across seven datasets, two of which contain biological and medical information that is subject to frequent unlearning requests.

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
- abstract@open-review(Poster): 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.

Federated Nearest Neighbor Machine Translation

- Yichao Du, Zhirui Zhang, Bingzhe Wu, Lemao Liu, Tong Xu, Enhong Chen
- abstract@open-review(Poster): 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 (kNN) 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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. Source code and configuration details will be all made publicly available.

Words are all you need? Language as an approximation for representational similarity

- Raja Marjieh, Pol Van Rijn, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, Nori Jacoby
- abstract@open-review(Poster): 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 modalities. 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.

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
- abstract@open-review(Poster): Pseudo labeling and consistency regularization approaches based on confidence thresholding have made great progress in semi-supervised learning (SSL). However, we argue that existing methods might fail to adopt suitable thresholds 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 achieve some implications on the relationship between the desirable threshold and model's learning status. Based on the analysis, we hence propose FreeMatch to define and 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 that encourages the model to produce diverse predictions during the early stages of training. Extensive experimental results 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.

Confidence Estimation Using Unlabeled Data

- Chen Li, Xiaoling Hu, Chao Chen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 lead 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.

SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-supervised Learning

- Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Marios Savvides, Jindong Wang, Bhiksha Raj, Xing Xie, Bernt Schiele

- abstract@open-review(Poster): The critical challenge of Semi-Supervised Learning (SSL) is how to effectively leverage 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
- abstract@open-review(Poster): 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.

Disentangling the Mechanisms Behind Implicit Regularization in SGD

- Zachary Novack, Simran Kaur, Tanya Marwah, Saurabh Garg, Zachary Chase Lipton
- abstract@open-review(Poster): 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.

Sequential Attention for Feature Selection

- Taisuke Yasuda, Mohammadhossein Bateni, Lin Chen, Matthew Fahrbach, Gang Fu, Vahab Mirrokni
- abstract@open-review(Poster): Feature selection is the problem of selecting a subset of features for a machine learning model that maximizes model quality subject to a resource budget constraint. For neural networks, prior methods, including those based on ℓ_1 regularization, attention, and stochastic gates, typically select all of the features in one evaluation round, ignoring the residual value of the features during selection (i.e., the marginal contribution of a feature conditioned on the previously selected features). 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 implementation of greedy forward selection and uses attention weights at each step as a proxy for feature importance. We provide theoretical insights into our Sequential Attention algorithm for linear regression models by showing that an adaptation to this setting is equivalent to the classical Orthogonal Matching Pursuit algorithm [PRK1993], and thus inherits all of its provable guarantees. Lastly, our theoretical and empirical analyses provide new explanations towards the effectiveness of attention and its connections to overparameterization, which might be of independent interest.

Improved Sample Complexity for Reward-free Reinforcement Learning under Low-rank MDPs

- Yuan Cheng, Ruiquan Huang, Yingbin Liang, Jing Yang
- abstract@open-review(Poster): 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. While reward-free RL has been well studied under the tabular setting with minimax optimal sample complexity being achieved, theoretical study of reward-free RL with complicated function approximation is still limited. In this paper we focus on reward-free RL under low-rank MDP models, which capture the representation learning in RL. We propose a new model-based algorithm, coined RAFFLE, and show that it can both find an ϵ-optimal policy and achieve an ϵ-accurate system identification via reward-free exploration, with a sample complexity of $\tilde{O}(\frac{H^3d^2K(d^2+K)}{\epsilon^2})$, where d, H and K respectively denote the representation dimension, episode horizon, and action space cardinality. This significantly improves the sample complexity of $\tilde{O}(\frac{H^{22}K^9d^7}{\epsilon^{10}})$ in Agarwal et al. (2020) for the same learning goals. We further provide a sample complexity lower bound of $\tilde{\Omega}(\frac{HdK}{\epsilon^2})$ that holds for any reward-free algorithm under low-rank MDPs, which matches our upper bound in the dependence on ϵ, as well as on K in the large d regime. Comparing this lower bound for low-rank MDPs with the upper bound for linear MDPs in Wang et al. (2020), it implies that reward-free RL under low-rank MDPs is strictly harder than linear MDPs. Finally, we complete our study by reusing RAFFLE to learn representation. We estimate the representation individually with only access to the learned transition kernels from RAFFLE and without interacting with true environment, and then theoretically characterize the closeness between the learned and the ground truth representation. The learned representation can be further used for few shot RL as in supervised learning (Du et al., 2021b).

Re-Imagen: Retrieval-Augmented Text-to-Image Generator

- Wenhu Chen, Hexiang Hu, Chitwan Saharia, William W. Cohen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 internalize experiences into a multi-task policy and rapidly solve new tasks.

Link Prediction with Non-Contrastive Learning

- William Shiao, Zhichun Guo, Tong Zhao, Evangelos E. Papalexakis, Yozen Liu, Neil Shah
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

A Theoretical Understanding of Vision Transformers: Learning, Generalization, and Sample Complexity

- Hongkang Li, Meng Wang, Sijia Liu, Pin-Yu Chen
- abstract@open-review(Poster): 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.

Contrastive Learning Can Find An Optimal Basis For Approximately View-Invariant Functions

- Daniel D. Johnson, Ayoub El Hanchi, Chris J. Maddison
- abstract@open-review(Poster): 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.

Provably Auditing Ordinary Least Squares in Low Dimensions

- Ankur Moitra, Dhruv Rohatgi
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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, time is assumed 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 edge classification and temporal link prediction.

Neural DAG Scheduling via One-Shot Priority Sampling

- Wonseok Jeon, Mukul Gagrani, Burak Bartan, Weiliang Will Zeng, Harris Teague, Piero Zappi, Christopher Lott
- abstract@open-review(Poster): 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.

Meta Temporal Point Processes

- Wonho Bae, Mohamed Osama Ahmed, Frederick Tung, Gabriel L. Oliveira
- abstract@open-review(Poster): 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.

Graph Neural Network-Inspired Kernels for Gaussian Processes in Semi-Supervised Learning

- Zehao Niu, Mihai Anitescu, Jie Chen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 invert`saccuracy-on-the-line" (Miller et al., 2021).

Diffusion Models for Causal Discovery via Topological Ordering

- Pedro Sanchez, Xiao Liu, Alison Q O'Neil, Sotirios A. Tsaftaris
- abstract@open-review(Poster): 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. In this case, the \emph{Hessian} of the data log-likelihood can be used for finding leaf nodes in a causal graph. Topological ordering approaches for causal discovery exploit this by performing graph discovery in two steps, first sequentially identifying nodes in reverse order of depth (\emph{topological ordering}), and secondly pruning the potential relations. This is more efficient since the search is performed over a permutation rather than a graph space. 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. Further, 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.

Scalable and Equivariant Spherical CNNs by Discrete-Continuous (DISCO) Convolutions

- Jeremy Ocampo, Matthew Alexander Price, Jason McEwen
- abstract@open-review(Poster): 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 not only exhibit $\text{SO}(3)$ rotational equivariance but also a form of asymptotic $\text{SO}(3)/\text{SO}(2)$ rotational equivariance, which is more desirable for many applications (where $\text{SO}(n)$ is the special orthogonal group representing rotations in n-dimensions). 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.

Weakly Supervised Explainable Phrasal Reasoning with Neural Fuzzy Logic

- Zijun Wu, Zi Xuan Zhang, Atharva Naik, Zhijian Mei, Mauajama Firdaus, Lili Mou
- abstract@open-review(Poster): 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 for 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.

DCI-ES: An Extended Disentanglement Framework with Connections to Identifiability

- Cian Eastwood, Andrei Liviu Nicolicioiu, Julius Von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, Bernhard Schölkopf
- abstract@open-review(Poster): 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, thus 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.

Faster federated optimization under second-order similarity

- Ahmed Khaled, Chi Jin
- abstract@open-review(Poster): 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.

Mutual Partial Label Learning with Competitive Label Noise

- Yan Yan, Yuhong Guo
- abstract@open-review(Poster): 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 noise labels. Most existing PLL methods assume the candidate noise labels are randomly chosen, which hardly holds in the real- world learning scenarios. In this paper, we consider a more realistic PLL scenario with competitive noise labels that are more difficult to distinguish from the true label than the random noise labels. 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 noise 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 the state-of-the-art performance for partial label learning.

Partial Label Unsupervised Domain Adaptation with Class-Prototype Alignment

- Yan Yan, Yuhong Guo
- abstract@open-review(Poster): 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. Most existing PLL approaches assume that both the training and test sets share the 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 named 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 domain discrepancy through inter- domain class prototype alignment. In addition, a teacher-student model based contrastive regularization is also deployed to enhance the prediction stability and hence improve the class prototypes in both domains for PLUDA. Comprehensive experimental results demonstrate that PAPLUDA achieves the state-of-the-art performance on the widely used benchmark datasets.

simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing

- Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Weiqi Luo
- abstract@open-review(Poster): 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 KCs. 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://tinyurl.com/5d62cdkt}.

Weighted Ensemble Self-Supervised Learning

- Yangjun Ruan, Saurabh Singh, Warren Richard Morningstar, Alexander A Alemi, Sergey Ioffe, Ian Fischer, Joshua V. Dillon
- abstract@open-review(Poster): 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.

Is a Caption Worth a Thousand Images? A Study on Representation Learning

- Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, Tatsunori Hashimoto

- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, Pradeep Kumar Ravikumar, Neil Y.C. Lin, Cho-Jui Hsieh
- abstract@open-review(Poster): 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 (potentially 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.

Constraining Representations Yields Models That Know What They Don't Know

- Joao Monteiro, Pau Rodriguez, Pierre-Andre Noel, Issam H. Laradji, David Vazquez
- abstract@open-review(Poster): 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.

An Extensible Multi-modal Multi-task Object Dataset with Materials

- Trevor Scott Standley, Ruohan Gao, Dawn Chen, Jiajun Wu, Silvio Savarese
- abstract@open-review(Poster): We present EMMa, an Extensible, Multimodal dataset of Amazon product listings that contains rich Material annotations. It contains more than 2 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 are annotated with one or more materials from this taxonomy. With the numerous data attributes available for each object, we develop a smart labeling framework to quickly add new binary labels to all objects with only hours of manual labeling effort, making the dataset extensible at scale. Each object attribute in our dataset can be included in either the model inputs or outputs, leading to combinatorial possibilities in task formulations. 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
- abstract@open-review(Poster): 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 target 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.

Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability

- Alex Damian, Eshaan Nichani, Jason D. Lee
- abstract@open-review(Poster): 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 et al. (2021) observed 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 edge 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.

TILP: Differentiable Learning of Temporal Logical Rules on Knowledge Graphs

- Siheng Xiong, Yuan Yang, Faramarz Fekri, James Clayton Kerce
- abstract@open-review(Poster): Compared with static knowledge graphs, temporal knowledge graphs (tKG), which can capture the evolution and change of information and knowledge, are more realistic and general. However, due to the complexity from introducing the notion of time, accurate link prediction based on explainable and comprehensible patterns is still a difficult problem. In this paper, we propose TILP, a differentiable framework for temporal logical rules learning. By designing constrained random walk mechanism and corresponding operators, we ensure the efficiency of our model. Furthermore, we discuss temporal features modelling in tKG, e.g., recurrence, temporal order, interval between pair of relations, and duration, and incorporate it into our learning process. TILP is compared with state-of-the-art baselines on two benchmark datasets and shows comparable performance. More importantly, we introduce some hard settings to test the robustness of different models, e.g., few training samples, biased data, and time shifting. In these cases, TILP works better than most state-of-the-art methods.

Open-Vocabulary Object Detection upon Frozen Vision and Language Models

- Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, Anelia Angelova
- abstract@open-review(Poster): 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.

Revisiting the Assumption of Latent Separability for Backdoor Defenses

- Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, Prateek Mittal
- abstract@open-review(Poster): 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.

Restricted Strong Convexity of Deep Learning Models with Smooth Activations

- Arindam Banerjee, Pedro Cisneros, Libin Zhu, Misha Belkin
- abstract@open-review(Poster): 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 thenear 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
- abstract@open-review(Poster): Temporal distributional shifts, with underlying dynamics changing over time, frequently occur in real-world time series, and pose a fundamental challenge for deep 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. To the best of our knowledge, this is the first time that Koopman theory is applied to real-world chaotic time series without known governing laws. We demonstrate that KNF achieves the superior performance compared to the alternatives, on multiple time series datasets that are shown to suffer from distribution shifts.

MetaGL: Evaluation-Free Selection of Graph Learning Models via Meta-Learning

- Namyong Park, Ryan A. Rossi, Nesreen Ahmed, Christos Faloutsos
- abstract@open-review(Poster): 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).

Minimum Description Length Control

- Ted Moskovitz, Ta-Chu Kao, Maneesh Sahani, Matthew Botvinick
- abstract@open-review(Poster): 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.

PerFedMask: Personalized Federated Learning with Optimized Masking Vectors

- Mehdi Setayesh, Xiaoxiao Li, Vincent W.S. Wong
- abstract@open-review(Poster): 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.

Variational Latent Branching Model for Off-Policy Evaluation

- Qitong Gao, Ge Gao, Min Chi, Miroslav Pajic
- abstract@open-review(Poster): 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.

Tuning Frequency Bias in Neural Network Training with Nonuniform Data

- Annan Yu, Yunan Yang, Alex Townsend
- abstract@open-review(Poster): 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.

Learning Multimodal Data Augmentation in Feature Space

- Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, Andrew Gordon Wilson
- abstract@open-review(Poster): 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.

BigVGAN: A Universal Neural Vocoder with Large-Scale Training

- Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, Sungroh Yoon
- abstract@open-review(Poster): 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 (OOD) 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 achieves the state-of-the-art performance for various scenarios, including new speakers, novel languages, unseen recording environments, singing voices, music and instrumental audio. Code and model will be released.

Achieving Sub-linear Regret in Infinite Horizon Average Reward Constrained MDP with Linear Function Approximation

- Arnob Ghosh, Xingyu Zhou, Ness Shroff
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

The Surprising Computational Power of Nondeterministic Stack RNNs

- Brian DuSell, David Chiang
- abstract@open-review(Poster): 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.

Agnostic Learning of General ReLU Activation Using Gradient Descent

- Pranjal Awasthi, Alex Tang, Aravindan Vijayaraghavan
- abstract@open-review(Poster): 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 descent, 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.

Learning Hyper Label Model for Programmatic Weak Supervision

- Renzhi Wu, Shen-En Chen, Jieyu Zhang, Xu Chu
- abstract@open-review(Poster): 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).

FedFA: Federated Feature Augmentation

- Tianfei Zhou, Ender Konukoglu
- abstract@open-review(Poster): 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 distinct perspective of federated feature augmentation. FedFA is based on a major insight that instance-level feature statistics (i.e., mean and standard deviation) represent a special type of ``features" that encodes domain-specific characteristics; hence, proper manipulations of the local feature statistics in the federation may beget novel domains, which can potentially alleviate local feature shift and benefit collaborative learning. Based on this insight, we model each feature statistic probabilistically via a Gaussian distribution, with the mean corresponding to the original statistic and the variance quantifying the augmentation scope, from which novel feature statistics can be drawn to fulfill augmentation. 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 demonstrate through extensive experiments that FedFA can significantly advance federated learning in diverse scenarios. The code will be released.

Offline Congestion Games: How Feedback Type Affects Data Coverage Requirement

- Haozhe Jiang, Qiwen Cui, Zhihan Xiong, Maryam Fazel, Simon Shaolei Du
- abstract@open-review(Poster): 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.

Does Decentralized Learning with Non-IID Unlabeled Data Benefit from Self Supervision?

- Lirui Wang, Kaiqing Zhang, Yunzhu Li, Yonglong Tian, Russ Tedrake
- abstract@open-review(Poster): 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 the 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 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.

Malign Overfitting: Interpolation and Invariance are Fundamentally at Odds

- Yoav Wald, Gal Yona, Uri Shalit, Yair Carmon
- abstract@open-review(Poster): 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 classifier (with nonzero 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 regarding the conflict between interpolation and invariance 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
- abstract@open-review(Poster): 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 moves 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.

SMART: Sentences as Basic Units for Text Evaluation

- Reinald Kim Amplayo, Peter J Liu, Yao Zhao, Shashi Narayan
- abstract@open-review(Poster): 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.

Tier Balancing: Towards Dynamic Fairness over Underlying Causal Factors

- Zeyu Tang, Yatong Chen, Yang Liu, Kun Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

AutoTransfer: AutoML with Knowledge Transfer - An Application to Graph Neural Networks

- Kaidi Cao, Jiaxuan You, Jiaju Liu, Jure Leskovec
- abstract@open-review(Poster): 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.

Temporal Dependencies in Feature Importance for Time Series Prediction

- Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, Maksims Volkovs
- abstract@open-review(Poster): 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.

Characterizing the spectrum of the NTK via a power series expansion

- Michael Murray, Hui Jin, Benjamin Bowman, Guido Montufar

- abstract@open-review(Poster): 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. 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 derive an explicit formula for the eigenvalues of the NTK, which shows faster decay in the NTK coefficients implies a faster decay in its spectrum. From this we recover existing results on eigenvalue asymptotics for ReLU networks and comment on how the activation function influences the RKHS. 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.

A critical look at evaluation of GNNs under heterophily: Are we really making progress?

- Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, Liudmila Prokhorenkova
- abstract@open-review(Poster): 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 considered that specialized methods are required to achieve strong performance on such graphs. Many GNN models for heterophilous graphs have recently been proposed in the literature. However, these models are typically evaluated on the same set of six heterophilous datasets. In this work, we challenge this evaluation setting. First, we show that the popular heterophilous benchmarks have serious drawbacks, the most significant being a large number of duplicate nodes in Squirrel and Chameleon datasets. We show that some models implicitly use this property of the datasets, and removing duplicate nodes strongly affects their performance. Then, we propose a set of heterophilous graphs of varying properties that we believe can serve as a better benchmark for testing the performance of GNNs under heterophily. We show that, at this moment, standard GNNs achieve competitive results on heterophilous graphs outperforming most of the specialized models.

A Non-monotonic Self-terminating Language Model

- Cheolhyoung Lee, Eugene Choi, Kyunghyun Cho
- abstract@open-review(Poster): 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.

Learning to Segment from Noisy Annotations: A Spatial Correction Approach

- Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, Chao Chen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 and speech 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.

MaskViT: Masked Visual Pre-Training for Video Prediction

- Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, Li Fei-Fei
- abstract@open-review(Poster): 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 approach, 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
- abstract@open-review(Poster): 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.

Continuous-time identification of dynamic state-space models by deep subspace encoding

- Gerben I. Beintema, Maarten Schoukens, Roland Tóth
- abstract@open-review(Poster): 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. This 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.

How to Train your HIPPO: State Space Models with Generalized Orthogonal Basis Projections

- Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, Christopher Re
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Layer Grafted Pre-training: Bridging Contrastive Learning And Masked Image Modeling For Better Representations

- Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, Zhangyang Wang
- abstract@open-review(Poster): 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. Our code will be released upon acceptance.

Discovering Latent Knowledge in Language Models Without Supervision

- Collin Burns, Haotian Ye, Dan Klein, Jacob Steinhardt
- abstract@open-review(Poster): 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.

Diffusion Adversarial Representation Learning for Self-supervised Vessel Segmentation

- Boah Kim, Yujin Oh, Jong Chul Ye
- abstract@open-review(Poster): 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 for vessel segmentation. In particular, for self-supervised vessel segmentation, DARL learns background image distribution 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 by one 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 methods in the vessel segmentation.

Noise Injection Node Regularization for Robust Learning

- Noam Itzhak Levi, Itay Mimouni Bloch, Marat Freytsis, Tomer Volansky
- abstract@open-review(Poster): 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.

Efficient Edge Inference by Selective Query

- Anil Kag, Igor Fedorov, Aditya Gangrade, Paul Whatmough, Venkatesh Saligrama
- abstract@open-review(Poster): 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.

Leveraging Incompatibility to Defend Against Backdoor Poisoning

- Charles Jin, Melinda Sun, Martin Rinard
- abstract@open-review(Poster): As deep learning datasets grow larger and less curated, backdoor data poisoning attacks, which inject malicious poisoned data into the training dataset, have drawn increasing attention in both academia and industry.

We identify an incompatibility property of the interaction of clean and poisoned data with the training algorithm, specifically that including poisoned data in the training dataset does not improve model accuracy on clean data and vice-versa. Leveraging this property, we develop an algorithm that iteratively refines subsets of the poisoned dataset to obtain subsets that concentrate around either clean or poisoned data. The result is a partition of the original dataset into disjoint subsets, for each of which we train a corresponding model. A voting algorithm over these models identifies the clean data within the larger poisoned dataset.

We empirically evaluate our approach and technique for image classification tasks over the GTSRB and CIFAR-10 datasets. The experimental results show that prior dirty-label and clean-label backdoor attacks in the literature produce poisoned datasets that exhibit behavior consistent with the incompatibility property. The results also show that our defense reduces the attack success rate below 1% on 134 out of 165 scenarios in this setting, with only a 2% drop in clean accuracy on CIFAR-10 (and negligible impact on GTSRB).

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
- abstract@open-review(Poster): Algorithms designed for single-agent reinforcement learning (RL) generally fail to converge to equilibria in two-player zero-sum (2p0s) games. On the other hand, game-theoretic algorithms for approximating Nash and regularized equilibria in 2p0s games are not typically competitive for RL and can be difficult to scale. As a result, algorithms for these two cases are generally developed and evaluated separately. In this work, we show that a single algorithm---a simple extension to mirror descent with proximal regularization that we call magnetic mirror descent (MMD)---can produce strong results in both settings, despite their fundamental differences. From a theoretical standpoint, we prove that MMD converges linearly to quantal response equilibria (i.e., entropy regularized Nash equilibria) in extensive-form games---this is the first time linear convergence has been proven for a first order solver. Moreover, applied as a tabular Nash equilibrium solver via self-play, we show empirically that MMD produces results competitive with CFR in both normal-form and extensive-form games---this is the first time that a standard RL algorithm has done so. Furthermore, for single-agent deep RL, on a small collection of Atari and Mujoco tasks, we show that MMD can produce results competitive with those of PPO. Lastly, for multi-agent deep RL, we show MMD can outperform NFSP in 3x3 Abrupt Dark Hex.

Pitfalls of Gaussians as a noise distribution in NCE

- Holden Lee, Chirag Pabbaraju, Anish Prasad Sevekari, Andrej Risteski
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Perfectly Secure Steganography Using Minimum Entropy Coupling

- Christian Schroeder de Witt, Samuel Sokota, J Zico Kolter, Jakob Nicolaus Foerster, Martin Strohmeier
- abstract@open-review(Poster): 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. Due to recent breakthroughs in approximate and iterative minimum entropy coupling techniques, 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 and WaveRNN as communication channels. We find that the minimum entropy coupling-based approach yields 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.

Calibrating Transformers via Sparse Gaussian Processes

- Wenlong Chen, Yingzhen Li
- abstract@open-review(Poster): 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 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.

Red PANDA: Disambiguating Anomaly Detection by Removing Nuisance Factors

- Niv Cohen, Jonathan Kahana, Yedid Hoshen
- abstract@open-review(Poster): Anomaly detection methods strive to discover patterns that differ from the norm in a meaningful way. This goal is ambiguous as different operators may find different attributes meaningful. A data point differing from the norm by an attribute such as pose, age, or gender, 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 learns 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.

Is Attention All That NeRF Needs?

- Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang
- abstract@open-review(Poster): We present Generalizable NeRF Transformer (GNT), a transformer-based architecture that reconstructs Neural Radiance Fields (NeRFs) and learns to renders 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 indicates that attention enables learning a physically-grounded rendering. Our results show the promise of transformers as a universal modelling tool for graphics.

STOCHASTIC NO-REGRET LEARNING FOR GENERAL GAMES WITH VARIANCE REDUCTION

- Yichi Zhou, Fang Kong, Shuai Li
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 raises concerns about the current practice of NAS and points to potential directions to develop effective countermeasures.

Generalization and Estimation Error Bounds for Model-based Neural Networks

- Avner Shultzman, Eyar Azar, Miguel R. D. Rodrigues, Yonina C. Eldar
- abstract@open-review(Poster): 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.

ChordMixer: A Scalable Neural Attention Model for Sequences with Different Length

- Ruslan Khalitov, Tong Yu, Lei Cheng, Zhirong Yang
- abstract@open-review(Poster): 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 signals 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
- abstract@open-review(Poster): The transferability of adversarial perturbations between image models has been extensively studied. In this case, an attack is generated from a known surrogate e.g., 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 image models such as Supervised Vision Transformers (ViTs), Self-supervised ViTs (e.g., DINO), and Vision-language models (e.g., CLIP) to black-box 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 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, e.g., divided space-time attention, 3D convolutions, or multi-view convolution networks for different data modalities. Image models are all we need to optimize for an effective surrogate and an adversarial attack to fool black-box models in a changing environment over time. Our code will be made public.

Static Prediction of Runtime Errors by Learning to Execute Programs with External Resource Descriptions

- David Bieber, Rishab Goel, Dan Zheng, Hugo Larochelle, Daniel Tarlow
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 46 650 OR-molecule pairs, this model outperforms standard GNN baselines by 30%. 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.

SGDA with shuffling: faster convergence for nonconvex-PŁ minimax optimization

- Hanseul Cho, Chulhee Yun
- abstract@open-review(Poster): 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-Łojasiewicz (PŁ) geometry. We analyze both simultaneous and alternating SGDA-RR for nonconvex-PŁ and primal-PŁ-PŁ objectives, and obtain convergence rates faster than with-replacement SGDA. Our rates also extend to mini-batch SGDA-RR, recovering known rates for full-batch gradient descent-ascent (GDA). Lastly, we present a comprehensive lower bound for two-time-scale GDA, which matches the full-batch rate for primal-PŁ-PŁ case.

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
- abstract@open-review(Poster): This paper presents MOAT, a family of neural networks that build on top of MObile convolution (i.e., inverted residual blocks) and ATtention. The Transformer architectures are gaining popularity in computer vision for their powerful attention mechanism to encode long-range interactions. However, their performance and generalization are worse than convolutional neural networks (ConvNets), especially in the low data regime. Transformers require a huge amount of training images to learn the right inductive biases for vision recognition, some of which (e.g., translation equivariance) are successfully built in ConvNets. The research community has thus attempted to combine the strengths from both architectures. Unlike the current works that simply 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% top-1 accuracy on ImageNet-1K 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 58.5% AP$^{\text{box}}$ with 110M 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. Code will be made publicly available.

Part-Based Models Improve Adversarial Robustness

- Chawin Sitawarin, Kornrapat Pongmala, Yizheng Chen, Nicholas Carlini, David Wagner
- abstract@open-review(Poster): 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 included in the supplementary material.

PGrad: Learning Principal Gradients For Domain Generalization

- Zhe Wang, Jake Grigsby, Yanjun Qi
- abstract@open-review(Poster): 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.

Extremely Simple Activation Shaping for Out-of-Distribution Detection

- Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, Rosanne Liu
- abstract@open-review(Poster): 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.

Statistical Guarantees for Consensus Clustering

- Zhixin Zhou, Gautam Dudeja, Arash A Amini
- abstract@open-review(Poster): 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 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 in N. Numerical experiments show the effectiveness of the proposed methods.

Expressive Monotonic Neural Networks

- Niklas Nolte, Ouail Kitouni, Mike Williams
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning Simultaneous Navigation and Construction in Grid Worlds

- Wenyu Han, Haoran Wu, Eisuke Hirota, Alexander Gao, Lerrel Pinto, Ludovic Righetti, Chen Feng
- abstract@open-review(Poster): 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.

Learning to CROSS exchange to solve min-max vehicle routing problems

- Minjun Kim, Junyoung Park, Jinkyoo Park
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 in the supplementary material.

Compositional Law Parsing with Latent Random Functions

- Fan Shi, Bin Li, Xiangyang Xue

- abstract@open-review(Poster): 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. In this paper, we propose a deep latent variable model for Compositional LAw Parsing (CLAP). CLAP achieves the human-like compositionality ability through an encoding-decoding architecture to represent concepts in the scene as latent variables, and further employ concept-specific random functions, instantiated with Neural Processes, in the latent space to capture the law on each concept. Our experimental results demonstrate that CLAP outperforms the compared baseline methods in multiple visual tasks including intuitive physics, abstract visual reasoning, and scene representation. In addition, CLAP can learn concept-specific laws in a scene without supervision and one can edit laws through modifying the corresponding latent random functions, validating its interpretability and manipulability.

LilNetX: Lightweight Networks with EXtreme Model Compression and Structured Sparsification

- Sharath Girish, Kamal Gupta, Saurabh Singh, Abhinav Shrivastava
- abstract@open-review(Poster): 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.

Mitigating Dataset Bias by Using Per-Sample Gradient

- Sumyeong Ahn, Seongyoon Kim, Se-Young Yun
- abstract@open-review(Poster): The performance of deep neural networks is strongly influenced by the training dataset setup. In particular, when attributes having 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 human. However, such methods require human costs. Recently, several studies have tried 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 PGD (Per-sample Gradient-based Debiasing). PGD comprises 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.

Efficient Model Updates for Approximate Unlearning of Graph-Structured Data

- Eli Chien, Chao Pan, Olgica Milenkovic
- abstract@open-review(Poster): 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.

AudioGen: Textually Guided Audio Generation

- Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, Yossi Adi
- abstract@open-review(Poster): 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.

Hebbian and Gradient-based Plasticity Enables Robust Memory and Rapid Learning in RNNs

- Yu Duan, Zhongfan Jia, Qian Li, Yi Zhong, Kaisheng Ma
- abstract@open-review(Poster): 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) with 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
- abstract@open-review(Poster): 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.

On the Data-Efficiency with Contrastive Image Transformation in Reinforcement Learning

- Sicong Liu, Xi Sheryl Zhang, Yushuo Li, Yifan Zhang, Jian Cheng
- abstract@open-review(Poster): Data-efficiency has always been an essential issue in pixel-based reinforcement learning (RL). As the agent not only learns the 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/Kamituna/CoIT.

Energy-based Out-of-Distribution Detection for Graph Neural Networks

- Qitian Wu, Yiting Chen, Chenxiao Yang, Junchi Yan
- abstract@open-review(Poster): 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 practical 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.

Quasi-optimal Learning with Continuous Treatments

- Yuhan Li, Wenzhuo Zhou, Ruoqing Zhu
- abstract@open-review(Poster): 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 \emph{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 Clients and Unbounded Losses

- Xiaolin Hu, Shaojie Li, Yong Liu
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 included in the supplementary.

Which Layer is Learning Faster? A Systematic Exploration of Layer-wise Convergence Rate for Deep Neural Networks

- Yixiong Chen, Alan Yuille, Zongwei Zhou
- abstract@open-review(Poster): 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.

A non-asymptotic analysis of oversmoothing in Graph Neural Networks

- Xinyi Wu, Zhengdao Chen, William Wei Wang, Ali Jadbabaie
- abstract@open-review(Poster): A central challenge of building more powerful Graph Neural Networks (GNNs) is the oversmoothing phenomenon, where increasing the network depth leads to homogeneous node representations and thus worse classification performance. 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-asymptotic 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) 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 may be exacerbated by the difficulty of optimizing deep GNN models.

Scaleformer: Iterative Multi-scale Refining Transformers for Time Series Forecasting

- Mohammad Amin Shabani, Amir H. Abdi, Lili Meng, Tristan Sylvain
- abstract@open-review(Poster): 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/Scaleformer/Scaleformer.

Liquid Structural State-Space Models

- Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, Daniela Rus
- abstract@open-review(Poster): 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 a large series of long-range sequence modeling benchmarks. In this paper, we show that we can improve further when the structural 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 structural state-space model, dubbed Liquid-S4, achieves the new state-of-the-art 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
- abstract@open-review(Poster): 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.

Ollivier-Ricci Curvature for Hypergraphs: A Unified Framework

- Corinna Coupette, Sebastian Dalleiger, Bastian Rieck
- abstract@open-review(Poster): 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, Ollivier-Ricci curvature measures differences between random walks via Wasserstein distances, thus grounding a geometric concept in ideas from probability 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.

Hard-Meta-Dataset++: Towards Understanding Few-Shot Performance on Difficult Tasks

- Samyadeep Basu, Megan Stanley, John F Bronskill, Soheil Feizi, Daniela Massiceti
- abstract@open-review(Poster): 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 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 our algorithm to extract difficult tasks from Meta-Dataset, 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-Meta-Dataset++, a new few-shot testing benchmark to promote the development of methods that are robust to even the most difficult tasks. We use Hard-Meta-Dataset++ 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 and Hard-Meta-Dataset++ will aid researchers in further understanding failure modes of few-shot classification models.

Compositional Semantic Parsing with Large Language Models

- Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, Denny Zhou
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 work, 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.

FaiREE: fair classification with finite-sample and distribution-free guarantee

- Puheng Li, James Zou, Linjun Zhang
- abstract@open-review(Poster): 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.

Exponential Generalization Bounds with Near-Optimal Rates for L_q-Stable Algorithms

- Xiaotong Yuan, Ping Li
- abstract@open-review(Poster): 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 tail generalization and 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 natural question we would like to address in this paper is whether it is possible to derive near-optimal exponential generalization 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. We demonstrate the power of our improved bounds by applying them 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Prompting GPT-3 To Be Reliable

- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, Lijuan Wang
- abstract@open-review(Poster): 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, existing research focuses on models' accuracy on standard benchmarks and largely ignores their reliability, which is crucial for avoiding catastrophic real-world harms. While reliability is a broad and vaguely defined term, this work decomposes reliability into four facets: generalizability, fairness, calibration, and factuality. We establish simple and effective prompts to demonstrate GPT-3's reliability in these four aspects: 1) generalize out-of-domain, 2) balance demographic distribution to reduce social biases, 3) calibrate language model probabilities, and 4) update the LLM's knowledge. We find that by employing appropriate prompts, GPT-3 outperforms smaller-scale supervised models by large margins on all these facets. We will release all pro- cessed datasets, evaluation scripts, and model predictions. Our findings not only shed new insights on the reliability of prompting LLMs, but more importantly, our prompting strategies can help practitioners more reliably use large language models like GPT-3.

Human alignment of neural network representations

- Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, Simon Kornblith
- abstract@open-review(Poster): 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 concept representations. Human representations are inferred from behavioral responses in an odd-one-out triplet or multi-arrangement task, where humans were required to make judgments about the similarities between different pairs of objects. 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. Learning a linear probe on behavioral responses for the odd-one-out task can substantially improve the alignment with human similarity judgments for the other tasks, without seeing that data. In addition, we find that some human concepts such as food and animals are well-represented in neural network representations 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.

Unbiased Stochastic Proximal Solver for Graph Neural Networks with Equilibrium States

- Mingjie Li, Yifei Wang, Yisen Wang, Zhouchen Lin
- abstract@open-review(Poster): 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 graph datasets show that our proposed solvers are more efficient and can achieve state-of-the-art performance.

DiGress: Discrete Denoising diffusion for graph generation

- Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, Pascal Frossard
- abstract@open-review(Poster): This work introduces DiGress, a discrete denoising diffusion model for generating graphs with categorical node and edge attributes. Our model defines a diffusion process that progressively edits a graph with noise (adding or removing edges, changing the categories), and a graph transformer network that learns to revert this process. With these two ingredients in place, we reduce distribution learning over graphs to a simple sequence of classification tasks. We further improve sample quality by proposing a new Markovian noise model that preserves the marginal distribution of node and edge types during diffusion, and by adding auxiliary graph-theoretic features derived from the noisy graph at each diffusion step. Finally, we propose a guidance procedure for conditioning the generation on graph-level features. Overall, DiGress achieves state-of-the-art performance on both molecular and non-molecular datasets, with up to 3x validity improvement on a dataset of planar graphs. In particular, it is the first model that scales to the large GuacaMol dataset containing 1.3M drug-like molecules without using a molecule-specific representation such as SMILES or fragments.

How to prepare your task head for finetuning

- Yi Ren, Shangmin Guo, Wonho Bae, Danica J. Sutherland
- abstract@open-review(Poster): 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.

Sequence to sequence text generation with diffusion models

- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, Lingpeng Kong
- abstract@open-review(Poster): 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 difficult due to the discrete nature of text. 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.

Policy Expansion for Bridging Offline-to-Online Reinforcement Learning

- Haichao Zhang, Wei Xu, Haonan Yu
- abstract@open-review(Poster): 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.

Mitigating Memorization of Noisy Labels via Regularization between Representations

- Hao Cheng, Zhaowei Zhu, Xing Sun, Yang Liu
- abstract@open-review(Poster): 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.

Graph Neural Networks are Inherently Good Generalizers: Insights by Bridging GNNs and MLPs

- Chenxiao Yang, Qitian Wu, Jiahua Wang, Junchi Yan
- abstract@open-review(Poster): 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 the conventional wisdom largely attributes the success of GNNs to their advanced expressivity for learning desired functions on nodes' ego-graphs, we conjecture that this is not the main cause of GNNs' superiority in node prediction tasks. This paper pinpoints the major source of GNNs' performance gain to their intrinsic generalization capabilities, by introducing an intermediate model class dubbed as P(ropagational)MLP, which is identical to standard MLP in training, and then adopt GNN's architecture in testing. Intriguingly, we observe that PMLPs consistently perform on par with (or even exceed) their GNN counterparts across ten benchmarks and different experimental settings, despite the fact that PMLPs share the same (trained) weights with poorly-performed MLP. This critical finding opens a door to a brand new perspective for understanding the power of GNNs, and allow bridging GNNs and MLPs for dissecting their generalization behaviors. As an initial step to analyze PMLP, we show its essential difference with MLP at infinite-width limit lies in the NTK feature map in the post-training stage. Moreover, though MLP and PMLP cannot extrapolate non-linear functions for extreme OOD data, PMLP has more freedom to generalize near the training support.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 sleep 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.

Improving Deep Policy Gradients with Value Function Search

- Enrico Marchesini, Christopher Amato
- abstract@open-review(Poster): Deep Policy Gradient 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. In this paper, we focus on improving value approximation and analyzing the effects on Deep Policy Gradient primitives such as value prediction, variance 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 Policy Gradient primitives results in improved sample efficiency and policies with higher returns using standard policy gradient methods on 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

A Learning Based Hypothesis Test for Harmful Covariate Shift

- Tom Ginsberg, Zhongyuan Liang, Rahul G Krishnan
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Non-parametric Outlier Synthesis

- Leitian Tao, Xuefeng Du, Jerry Zhu, Yixuan Li
- abstract@open-review(Poster): 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.

Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks

- Namjoon Suh, Tian-Yi Zhou, Xiaoming Huo
- abstract@open-review(Poster): 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.

Learning Adversarial Linear Mixture Markov Decision Processes with Bandit Feedback and Unknown Transition

- Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, Shuai Li
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 Neural Mean Embedding Approach for Back-door and Front-door Adjustment

- Liyuan Xu, Arthur Gretton
- abstract@open-review(Poster): 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.

TranSpeech: Speech-to-Speech Translation With Bilateral Perturbation

- Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, Zhou Zhao
- abstract@open-review(Poster): Direct speech-to-speech translation (S2ST) with discrete units leverages recent progress in speech representation learning, where a sequence of discrete representations derived in a self-supervised manner, are predicted from the model and passed to a vocoder for speech synthesis, 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 of 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/.

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
- abstract@open-review(Poster): 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.

Jointly Learning Visual and Auditory Speech Representations from Raw Data

- Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Maja Pantic
- abstract@open-review(Poster): 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 were jointly trained. Notably, combining RAVEn pre-training with self-training using only 30 hours of labelled data achieves competitive performance for visual speech recognition on LRS3, surpassing all self-supervised methods and a recent fully-supervised method trained on 90,000 hours of non-public labelled data. At the same time, we are on par with the state-of-the-art for auditory speech recognition on LRS3. 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 will be made public.

Diminishing Return of Value Expansion Methods in Model-Based Reinforcement Learning

- Daniel Palenicek, Michael Lutter, Joao Carvalho, Jan Peters
- abstract@open-review(Poster): Model-based reinforcement learning is an approach to increase sample efficiency. However, the accuracy of the dynamics models and the resulting compounding error over trajectories are commonly regarded as a limitation of model-based approaches. A natural question to ask is: How much more sample efficiency can be gained by improving the learned dynamics models? Specifically, this paper addresses the value expansion class of model-based approaches. Our empirical study shows that expanding the value function for the critic or actor update increases sample efficiency, but the gain in improvement decreases with each added expansion step. Therefore, longer horizons yield diminishing returns in terms of sample efficiency. In an extensive experimental comparison that uses the oracle dynamics model to avoid compounding model error, we show that short horizons are sufficient to obtain the lowest sample complexity for the given tasks. For long horizons, the improvements are marginal or can even decrease learning performance despite using the oracle dynamics model. Model-free counterparts, which use off-policy trajectories from a replay buffer and introduce no computational overhead, often show on-par performance and pose as a strong baseline. Finally, as we observe the same issues with both oracle and learned models, we conclude that the limitation of model-based value expansion methods is not so much the model accuracy of the learned models.

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
- abstract@open-review(Poster): 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 propose 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 will release the code and model to facilitate future research.

Equivariant Energy-Guided SDE for Inverse Molecular Design

- Fan Bao, Min Zhao, Zhongkai Hao, Peiyao Li, Chongxuan Li, Jun Zhu
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 proper pretraining and concurrent cross-task online fine-tuning, 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.

A Simple Yet Powerful Deep Active Learning With Snapshots Ensembles

- Seohyeon Jung, Sanghyun Kim, Juho Lee
- abstract@open-review(Poster): 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.

Decoupled Training for Long-Tailed Classification With Stochastic Representations

- Giung Nam, Sunguk Jang, Juho Lee
- abstract@open-review(Poster): 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. Experiments on ImageNet-LT and iNaturalist-2018 benchmarks show that our proposed method improves upon previous methods both in terms of prediction accuracy and uncertainty estimation.

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
- abstract@open-review(Poster): 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.

Benchmarking Constraint Inference in Inverse Reinforcement Learning

- Guiliang Liu, Yudong Luo, Ashish Gaurav, Kasra Rezaee, Pascal Poupart
- abstract@open-review(Poster): 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 followed by expert agents (e.g., humans) are often hard to specify mathematically and unknown to the RL agents. To tackle these issues, Inverse Constrained Reinforcement Learning (ICRL) considers the formalism of Constrained Markov Decision Processes (CMDPs) and estimates constraints from expert demonstrations by learning a constraint function. As an emerging research topic, ICRL does not have common benchmarks, and previous works tested their algorithms with hand-crafted environments (e.g., grid worlds). In this paper, we construct an ICRL benchmark in the context of two major application domains: robot control and autonomous driving. For each environment, we design relevant constraints, generate the corresponding expert trajectories, and empirically justify the importance of these constraints. To recover the constraints from expert demonstrations, previous ICRL methods typically learn a deterministic constraint function, which might dismiss the true constraint during training. We tackle this issue by proposing a variational Bayesian approach to model the posterior distribution of candidate constraints. Empirical evaluation shows this method outperforms other baselines in terms of collecting rewards and satisfying constraints. The benchmark, including the instructions for reproducing ICRL algorithms, is available at~{\it temporally hidden due to the anonymous policy}.

Memory Gym: Partially Observable Challenges to Memory-Based Agents

- Marco Pleines, Matthias Pallasch, Frank Zimmer, Mike Preuss
- abstract@open-review(Poster): 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 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. Several commonly used related environments do not share those qualities. 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 leverage recurrent model architectures as their memory.

Discovering Policies with DOMiNO: Diversity Optimization Maintaining Near Optimality

- Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, Satinder Singh
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 overall 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 a novel algorithmic technique, Priority Refresh, to stabilize massively parallel model-based training. SpeedyZero maintains on-par sample efficiency compared with EfficientZero while achieving a 20X speedup in wall-clock time, leading to human-level performances on the Atari benchmark within 30 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.

Neural Architecture Design and Robustness: A Dataset

- Steffen Jung, Jovita Lukasik, Margret Keuper
- abstract@open-review(Poster): 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 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 $6\,466$ 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 $0.20\%-0.41\%$.

Does Deep Learning Learn to Abstract? A Systematic Probing Framework

- Shengnan An, Zeqi Lin, Bei Chen, Qiang Fu, Nanning Zheng, Jian-Guang Lou
- abstract@open-review(Poster): 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) PLMs exhibit better abstraction capability with larger model sizes, larger data scales, and higher diversity in data.

Improving Out-of-distribution Generalization with Indirection Representations

- Kha Pham, Hung Le, Man Ngo, Truyen Tran
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Guided diffusion is a technique for conditioning the output of a diffusion model at sampling time without retraining the network for each specific task. One drawback of diffusion models, however, is their 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 high-order numerical methods are unsuitable for the conditional function. Our proposed method can re-utilize 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.

Batch Multivalid Conformal Prediction

- Christopher Jung, Georgy Noarov, Ramya Ramalingam, Aaron Roth
- abstract@open-review(Poster): 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 level $1-\alpha$ 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.

Accurate Bayesian Meta-Learning by Accurate Task Posterior Inference

- Michael Volpp, Philipp Dahlinger, Philipp Becker, Christian Daniel, Gerhard Neumann
- abstract@open-review(Poster): 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, 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 (i) GMM-NP yields tighter evidence lower bounds, increasing the efficiency of marginal likelihood optimization, leading to (ii) improved epistemic uncertainty estimation and accuracy, (iii) without any complex architectural modifications, resulting in a powerful, yet (iv) conceptually simple BML model. GMM-NP outperforms the state of the art on a range of challenging experiments, which highlight 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
- abstract@open-review(Poster): 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%.

Context-enriched molecule representations improve few-shot drug discovery

- Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, Günter Klambauer
- abstract@open-review(Poster): 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.

Test-Time Adaptation via Self-Training with Nearest Neighbor Information

- Minguk Jang, Sae-Young Chung, Hye Won Chung
- abstract@open-review(Poster): 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. Our experiments on two standard benchmark tasks, domain generalization and image corruption, show that TAST outperforms the state-of-the-art TTA methods.

Accurate Neural Training with 4-bit Matrix Multiplications at Standard Formats

- Brian Chmiel, Ron Banner, Elad Hoffer, Hilla Ben-Yaacov, Daniel Soudry
- abstract@open-review(Poster): 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.

Unsupervised Manifold Alignment with Joint Multidimensional Scaling

- Dexiong Chen, Bowen Fan, Carlos Oliver, Karsten Borgwardt
- abstract@open-review(Poster): 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.

Simple and Scalable Nearest Neighbor Machine Translation

- Yuhan Dai, Zhirui Zhang, Qiuzhi Liu, Qu Cui, Weihua Li, Yichao Du, Tong Xu
- abstract@open-review(Poster): kNN-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 (kNN) retrieval to achieve domain adaptation without retraining. Despite being conceptually attractive, kNN-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 kNN-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 kNN-MT, based on which we further introduce a distance-aware adapter to adaptively incorporate the kNN 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 kNN-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
- abstract@open-review(Poster): 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.

Dynamic Update-to-Data Ratio: Minimizing World Model Overfitting

- Nicolai Dorka, Tim Welschehold, Wolfram Burgard
- abstract@open-review(Poster): 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.

A Universal 3D Molecular Representation Learning Framework

- Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, Guolin Ke
- abstract@open-review(Poster): 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 that significantly enlarges the representation ability and application scope of MRL schemes. The proposed framework contains two pretrained models with the same SE(3)-equivariant transformer architecture: a molecular model pretrained by 209M molecular conformations; a pocket model pretrained by 3M candidate protein pocket data. Besides, the framework contains several finetuning strategies to apply the pretrained models to various downstream tasks. By properly incorporating 3D information, the framework outperforms SOTA in 14/15 molecular property prediction tasks. Moreover, the framework achieves superior performance in 3D spatial tasks, including protein-ligand binding pose prediction, molecular conformation generation, etc. The code, model, and data will be made publicly available.

Learning with Auxiliary Activation for Memory-Efficient Training

- Sunghyeon Woo, Dongsuk Jeon
- abstract@open-review(Poster): 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.

Massively Scaling Heteroscedastic Classifiers

- Mark Collier, Rodolphe Jenatton, Basil Mustafa, Neil Houlsby, Jesse Berent, Effrosyni Kokiopoulou
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Finding the global semantic representation in GAN through Fréchet Mean

- Jaewoong Choi, Geonho Hwang, Hyunsoo Cho, Myungjoo Kang
- abstract@open-review(Poster): 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 generating the same semantic subspace as the previous method.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 explicitly 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.

Blurring Diffusion Models

- Emiel Hoogeboom, Tim Salimans
- abstract@open-review(Poster): 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 heat dissipation, which we call Blurring Diffusion Models.

Hyperbolic Self-paced Learning for Self-supervised Skeleton-based Action Representations

- Luca Franco, Paolo Mandica, Bharti Munjal, Fabio Galasso
- abstract@open-review(Poster): 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 skeleton-based 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 enclosed in the submission and will be released.

Efficient Offline Policy Optimization with a Learned Model

- Zichen Liu, Siyi Li, Wee Sun Lee, Shuicheng YAN, Zhongwen Xu
- abstract@open-review(Poster): 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.

New Insights for the Stability-Plasticity Dilemma in Online Continual Learning

- Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, Sungroh Yoon
- abstract@open-review(Poster): 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. We will release our code upon acceptance.

MixPro: Data Augmentation with MaskMix and Progressive Attention Labeling for Vision Transformer

- Qihao Zhao, Yangyu Huang, Wei Hu, Fan Zhang, Jun Liu
- abstract@open-review(Poster): 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: Disentangling Shape, Pose and Appearance through 3D Morphable Image and Geometry Generation

- Eric-Tuan Le, Edward Bartrum, Iasonas Kokkinos
- abstract@open-review(Poster): We introduce StyleMorph, a 3D generative model that relies on the 3D morphable model paradigm to disentangle shape, pose, object and scene texture for high quality image synthesis. We represent 3D shape variability through 3D deformation fields with respect to a canonical object template. Both the deformations and the template are expressed as implicit networks and learned in an unsupervised manner only from 2D image supervision. We connect 3D morphable modelling with deferred neural rendering by performing an implicit surface rendering of “Template Object Coordinates” (TOCS), thereby constructing a purely geometric, deformation-equivariant 2D signal that reflects the compounded geometric effects of non-rigid shape, pose, and perspective projection. We use TOCS maps in tandem with object and background appearance codes to condition a StyleGAN-based deferred neural rendering (DNR) network for high-resolution image synthesis. We show competitive photorrealistic image synthesis results on 4 datasets (FFHQ faces, AFHQ Cats, Dogs, Wild), while achieving the joint disentanglement of shape, pose, object and scene texture.

Searching Lottery Tickets in Graph Neural Networks: A Dual Perspective

- Kun Wang, Yuxuan Liang, Pengkun Wang, Xu Wang, Pengfei Gu, Junfeng Fang, Yang Wang
- abstract@open-review(Poster): 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, Long Chen, Jun Xiao
- abstract@open-review(Poster): 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 different 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.

Unsupervised visualization of image datasets using contrastive learning

- Niklas Böhm, Philipp Berens, Dmitry Kobak
- abstract@open-review(Poster): 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.

PowerQuant: Automorphism Search for Non-Uniform Quantization

- Edouard YVINEC, Arnaud Dapogny, Matthieu Cord, Kevin Bailly
- abstract@open-review(Poster): 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.

BAYES RISK CTC: CONTROLLABLE CTC ALIGNMENT IN SEQUENCE-TO-SEQUENCE TASKS

- Jinchuan Tian, Brian Yan, Jianwei Yu, CHAO WENG, Dong Yu, Shinji Watanabe
- abstract@open-review(Poster): 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.

Bayesian semi-supervised learning with a principled likelihood from a generative model of data curation

- Stoil Krasimirov Ganev, Laurence Aitchison
- abstract@open-review(Poster): 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 Bayesian 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 an 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.

E3Bind: An End-to-End Equivariant Network for Protein-Ligand Docking

- Yangtian Zhang, Huiyu Cai, Chence Shi, Jian Tang
- abstract@open-review(Poster): 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.

Exact Group Fairness Regularization via Classwise Robust Optimization

- Sangwon Jung, Taeeon Park, Sanghyuk Chun, Taesup Moon
- abstract@open-review(Poster): Existing group fairness-aware training methods typically employ some heuristics, such as re-weighting underrepresented groups based on some rules or using approximated surrogates for the exact fairness metrics as regularization terms, which result in models with sub-optimal accuracy-fairness trade-offs. The reason for using such heuristics is that the fairness metrics are usually non-differentiable or non-convex, and exactly incorporating those metrics in a tractable learning objective is challenging. To that end, we propose a principled method that indeed can incorporate an $\textit{exact}$ form of a well-justified group fairness metric, Difference of Conditional Accuracy (DCA), as a regularizer using a $\textit{classwise}$ distributionally robust optimization (DRO) framework. Namely, we first show that the DCA is equivalent (up to a constant) to the average (over the classes) of the roots of the $\textit{variances}$ of group losses, then employ the Group DRO formulation for each class $\textit{separately}$ to convert the non-differentiable DCA (or variance) regularized group-balanced empirical risk minimization to a more tractable minimax optimization. We further develop an efficient iterative optimization algorithm and show that our resulting method, dubbed as FairDRO, makes an interesting connection between the re-weighting based and regularization-based fairness-aware learning. Our experiments show that FairDRO is scalable, easily adaptable to diverse applications, and consistently improves the group fairness on several benchmark datasets in terms of the accuracy-fairness trade-off, compared to recent state-of-the-art baselines.

Are More Layers Beneficial to Graph Transformers?

- Haiteng Zhao, Shuming Ma, Dongdong Zhang, Zhi-Hong Deng, Furu Wei
- abstract@open-review(Poster): 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.

Simplicial Hopfield networks

- Thomas F Burns, Tomoki Fukai
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Classically Approximating Variational Quantum Machine Learning with Random Fourier Features

- Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, Elham Kashefi

- abstract@open-review(Poster): 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. Precisely, 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
- abstract@open-review(Poster): 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 regime, 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 which 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; 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.

A Differential Geometric View and Explainability of GNN on Evolving Graphs

- Yazheng Liu, Xi Zhang, Sihong Xie
- abstract@open-review(Poster): 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 approximate 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 Cheng
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning

- Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, Yoon Kim
- abstract@open-review(Poster): 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 the efficient adaptation of 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 in task-specific prompt vectors to improve performance on target downstream tasks. In this paper, we propose multitask prompt tuning (MPT), which first learns a single transferable prompt by decomposing and 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 21 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.

Better Generative Replay for Continual Federated Learning

- Daiqing Qi, Handong Zhao, Sheng Li
- abstract@open-review(Poster): 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 sce- nario 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.

Generative Modelling with Inverse Heat Dissipation

- Severi Rissanen, Markus Heinonen, Arno Solin
- abstract@open-review(Poster): 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 model that generates images through iteratively inverting the heat equation, a PDE that locally erases fine-scale information when run over the 2D plane of the image. We interpret a noise-relaxed solution of the forward heat equation as a variational approximation in a diffusion-like 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 and data efficiency. Spectral analysis on natural images highlights connections to diffusion models and reveals implicit inductive biases in them.

Self-supervision through Random Segments with Autoregressive Coding (RandSAC)

- Tianyu Hua, Yonglong Tian, Sucheng Ren, Michalis Raptis, Hang Zhao, Leonid Sigal
- abstract@open-review(Poster): 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.

TRANSFORMER-PATCHER: ONE MISTAKE WORTH ONE NEURON

- Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, Zhang Xiong
- abstract@open-review(Poster): 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).

Sharper Bounds for Uniformly Stable Algorithms with Stationary φ-mixing Process

- Shi Fu, Yunwen Lei, Qiong Cao, Xinmei Tian, Dacheng Tao
- abstract@open-review(Poster): Generalization analysis of learning algorithms often builds on a critical assumption that training examples are independently and identically distributed (i.i.d.), 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 φ-mixing data, where the dependency between observations weakens over time. We show uniformly stable algorithms guarantee high-probability generalization bounds, which significantly improves the state of the art by a factor of $\sqrt{n}$ with n being 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 φ-mixing data. Our analysis builds on a novel moment bound for weakly-dependent random variables on a mixing sequence and a novel error decomposition of generalization error.

On the Edge of Benign Overfitting: Label Noise and Overparameterization Level

- Kaiyue Wen, Jiaye Teng, Jingzhao Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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. Experiments on various predictive inference tasks corroborate the efficacy of our method.

Recon: Reducing Conflicting Gradients From the Root For Multi-Task Learning

- Guangyuan SHI, Qimai Li, Wenlong Zhang, Jiaxin Chen, Xiao-Ming Wu
- abstract@open-review(Poster): 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 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 set them task-specific. Our experiments show that with only a slight increase in model parameters, such a simple approach can effectively reduce the occurrence of conflicting gradients in the remaining shared layers and achieve better performance. We demonstrate the generality of our approach by combining it with state-of-the-art approaches including gradient manipulation methods and branched architecture search methods. Comprehensive experiments on various benchmarks show that our approach can substantially improve their performance.

Measure the Predictive Heterogeneity

- Jiashuo Liu, Jiayun Wu, Renjie Pi, Renzhe Xu, Xingxuan Zhang, Bo Li, Peng Cui

- abstract@open-review(Poster): 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 \emph{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.

Time to augment visual self-supervised learning

- Arthur Aubret, Markus R. Ernst, Céline Teulière, Jochen Triesch
- abstract@open-review(Poster): 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. These 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. 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 for learning object categories. Our results show that time-based augmentations achieve 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 vital for learning to discard background-related information. Overall, we conclude that time-based augmentations can greatly improve contrastive learning, narrowing the gap between artificial and biological vision systems.

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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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, the proposed algorithm can outperform the benchmark MORL methods on a variety of MORL benchmark tasks.

Variance-Aware Sparse Linear Bandits

- Yan Dai, Ruosong Wang, Simon Shaolei Du
- abstract@open-review(Poster): 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.

CircNet: Meshing 3D Point Clouds with Circumcenter Detection

- Huan Lei, Ruitao Leng, Liang Zheng, Hongdong Li
- abstract@open-review(Poster): 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 this link.

In-sample Actor Critic for Offline Reinforcement Learning

- Hongchang Zhang, Yixiu Mao, Boyuan Wang, Shuncheng He, Yi Xu, Xiangyang Ji
- abstract@open-review(Poster): 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 performs 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

- abstract@open-review(Poster): Understanding the interaction between multiple agents is crucial for realistic and plausible vehicle trajectory prediction. Accordingly, existing methods tried to model and predict the interaction using observed past trajectories of agents with pooling, attention, or graph-based methods. However, we observed that they easily fail under complex road structures. It is because they do not explicitly utilize the map information for predicting the relationship, and they only model the relationship between vehicles in a deterministic manner, not a stochastic manner. In this paper, we propose a new method to formulate a stochastic future relationship among agents using lane structure. Our method first predicts a probability of lane-level waypoint occupancy of vehicles. Then we utilizes the temporal probability of passing the same lanes to learn the interaction between agents. In addition, we model the interaction using probabilistic distribution. This distribution is trained by posterior distribution of interaction from GT future trajectory. We validate our method on popular trajectory prediction datasets: nuScenes and Argoverse. The code will be available in public upon acceptance.

LMSeg: Language-guided Multi-dataset Segmentation

- Qiang Zhou, Yuang Liu, Chaohui Yu, Jingliang Li, Zhibin Wang, Fan Wang
- abstract@open-review(Poster): 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 proper- ties 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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, while 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
- abstract@open-review(Poster): 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.

Toward Adversarial Training on Contextualized Language Representation

- Hongqiu Wu, Yongxiang Liu, Hanwen Shi, hai zhao, Min Zhang
- abstract@open-review(Poster): Beyond the success story of adversarial training (AT) in the recent text domain on top of pre-trained language models (PrLMs), our empirical results showcase that current AT can appear mediocre or even harmful on certain tasks, e.g. reading comprehension and commonsense reasoning. This paper investigates AT from the perspective of contextualized language representation. We find that the gain from AT does not derive from increasing the training risk, but from deviating the language representation. The fact is that the current AT attack is better at fooling the decoder (i.e. the classifier), but can be trivial to the encoder. Based on the observations, we propose simple yet effective \textit{Contextualized representation-Adversarial Training} (CreAT), in which the attack is explicitly optimized to deviate the contextualized representation and obtains the global worst-case adversarial examples. CreAT is proven to be all-powerful compared to AT, with performance gain covering a wider range of downstream tasks. We apply CreAT to language pre-training. Our CreAT-empowered DeBERTa outperforms naive DeBERTa by a large margin, achieving the new state-of-the-art performances on a wide range of challenging benchmarks, e.g. AdvGLUE (59.1 $ \rightarrow $ 61.1), HellaSWAG (93.0 $ \rightarrow $ 94.9), ANLI (68.1 $ \rightarrow $ 69.3), PAWS (50.3 $ \rightarrow $ 54.5).

Gromov-Wasserstein Autoencoders

- Nao Nakagawa, Ren Togo, Takahiro Ogawa, Miki Haseyama
- abstract@open-review(Poster): 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.

Optimal Activation Functions for the Random Features Regression Model

- Jianxin Wang, José Bento
- abstract@open-review(Poster): 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.

Unsupervised Object-Centric Learning with Bi-level Optimized Query Slot Attention

- Baoxiong Jia, Yu Liu, Siyuan Huang
- abstract@open-review(Poster): 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 complex natural scenes. In this paper, we propose to address these issues by (1) initializing Slot-Attention modules with learnable queries and (2) optimizing the model with bi-level optimization. With simple code adjustments on the vanilla Slot-Attention, our model, BO-QSA, achieves state-of-the-art results on both synthetic and complex real-world datasets in unsupervised image segmentation and reconstruction, outperforming previous baselines by a large margin (~10%). We provide thorough ablative studies to validate the necessity and effectiveness of our design. Additionally, our model exhibits excellent potential for concept binding and zero-shot learning. We hope our effort could provide a single home for the design and learning of slot-based models and pave the way for more challenging tasks in object-centric learning.

Learning an Invertible Output Mapping Can Mitigate Simplicity Bias in Neural Networks

- Sravanti Addepalli, Anshul Nasery, Venkatesh Babu Radhakrishnan, Praneeth Netrapalli, Prateek Jain
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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. Codes and visualization videos are released on our homepage.

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
- abstract@open-review(Poster): 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 information 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
- abstract@open-review(Poster): 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 to Improving Exploration

- Dinghuai Zhang, Aaron Courville, Yoshua Bengio, Qinqing Zheng, Amy Zhang, Ricky T. Q. Chen
- abstract@open-review(Poster): 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.

Generalizing and Decoupling Neural Collapse via Hyperspherical Uniformity Gap

- Weiyang Liu, Longhui Yu, Adrian Weller, Bernhard Schölkopf
- abstract@open-review(Poster): 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. Inspired by how NC characterizes the training target of neural networks, we decouple NC into two objectives: minimal intra-class variability and maximal inter-class separability. We then introduce the concept of 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 generic 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 NC, but also decouples NC 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, calibration 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
- abstract@open-review(Poster): 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.

Empirical Study of Pre-training a Backbone for 3D Human Pose and Shape Estimation

- Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, Kyoung Mu Lee
- abstract@open-review(Poster): We empirically study unexplored, yet must-know baselines of pre-training a backbone for 3D human pose and shape estimation (3DHPSE). Recently, a few self-supervised representation learning (SSL) methods have been reported to outperform the ImageNet classification pre-training for vision tasks such as object detection. However, its effects on 3DHPSE are open to question, whose target is fixed to a single class, the human. In this regard, we inspect the effectiveness of SSL on 3DHPSE and investigate two other pre-training approaches that have received relatively less attention. They are 2D annotation-based pre-training and synthetic data pre-training. Similar to the motivation of SSL to benefit from unlabeled data, they have potential advantages to exploit data with less data collection cost compared with real 3D data. 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. In the semi-supervised setting, the improvement increases up to 8.2%, while SSL decreases accuracy by 10.7%, and synthetic data pre-training shows 0.2% decreased accuracy compared with the classification pre-training. Our observations would make the community carefully think about the current SSL-based pre-training trend for 3DHPSE and diversify research on pre-training approaches.

Learned Index with Dynamic ϵ

- Daoyuan Chen, Wuchao Li, Yaliang Li, Bolin Ding, Kai Zeng, Defu Lian, Jingren Zhou
- abstract@open-review(Poster): 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.

Breaking the Curse of Dimensionality in Multiagent State Space: A Unified Agent Permutation Framework

- Jianye HAO, Xiaotian Hao, Hangyu Mao, Weixun Wang, Yaodong Yang, Dong Li, YAN ZHENG, Zhen Wang
- abstract@open-review(Poster): 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 separate 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).

Wav2Tok: Deep Sequence Tokenizer for Audio Retrieval

- Adhiraj Banerjee, Vipul Arora
- abstract@open-review(Poster): 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 discrete 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.

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
- abstract@open-review(Poster): 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 animating static portraits and view-consistent motion editing. Code and models are released at the project page https://showlab.github.io/pv3d.

Characterizing the Influence of Graph Elements

- Zizhang Chen, Peizhao Li, Hongfu Liu, Pengyu Hong
- abstract@open-review(Poster): 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.

SWIFT: Rapid Decentralized Federated Learning via Wait-Free Model Communication

- Marco Bornstein, Tahseen Rabbani, Evan Z Wang, Amrit Bedi, Furong Huang
- abstract@open-review(Poster): 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 time 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
- abstract@open-review(Poster): 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.

Prototypical Calibration for Few-shot Learning of Language Models

- Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, Furu Wei
- abstract@open-review(Poster): 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 Single Noisy Images

- Heewon Kim, Kyoung Mu Lee
- abstract@open-review(Poster): 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 (\textit{e.g.}, DnCNN, ResNet-style network) outperforming previous noise-synthesis-based and self-supervision-based denoisers in real datasets.

Hierarchical Protein Representations via Complete 3D Graph Networks

- Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, Shuiwang Ji
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): One-shot coreset selection aims to select a subset of the training data, given a pruning rate, that can achieve high accuracy for models that are subsequently trained only with that subset. State-of-the-art coreset selection methods typically assign an importance score to each example and select the most important examples to form a coreset. These methods perform well at low pruning rates; but at high pruning rates, they have been found to suffer a catastrophic accuracy drop, performing worse than even random coreset selection. In this paper, we explore the reasons for this accuracy drop both theoretically and empirically. We extend previous theoretical results on the bound for model loss in terms of coverage provided by the coreset. Inspired by theoretical results, we propose a novel coverage-based metric and, based on the metric, find that coresets selected by importance-based coreset methods at high pruning rates can be expected to perform poorly compared to random coresets because of worse data coverage. We then propose a new coreset selection method, Coreset-centric Coreset Selection (CCS), where we jointly consider overall data coverage based on the proposed metric as well as importance of each example. We evaluate CCS on four datasets and show that they achieve significantly better accuracy than state-of-the-art coreset selection methods as well as random sampling under high pruning rates, and comparable performance at low pruning rates. For example, CCS achieves 7.04% better accuracy than random sampling and at least 20.16% better than popular importance-based selection methods on CIFAR10 with a 90% pruning rate.

ILA-DA: Improving Transferability of Intermediate Level Attack with Data Augmentation

- Chiu Wai Yan, Tsz-Him Cheung, Dit-Yan Yeung
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): The creation of contrastive vision-language models has traditionally required aligning a vision model with a language model by updating all of their parameters through gradient descent. It is not known if contrastive vision-language models (e.g. CLIP) can be created by a small number of parameter updates to already-trained language and vision models. The literature describes techniques that can create vision-language models by updating a small number of parameters 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. We show evidence of an intriguing asymmetry in the vision and language models, and how it affects alignment. 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 current full-model training paradigm for common use cases.

3EF: Class-Incremental Learning via Efficient Energy-Based Expansion and Fusion

- Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Yatao Bian, Han-Jia Ye, De-Chuan Zhan, Peilin Zhao
- abstract@open-review(Poster): Neural networks suffer from catastrophic forgetting when sequentially learning tasks phase-by-phase, making them inapplicable in dynamically renewal systems. Class-incremental learning (CIL) aims to enable neural networks to learn different categories at multi-stages. Recently, dynamic-structure-based methods achieves remarkable performance. However, these methods train all modules in a coupled manner and do not consider possible conflicts among modules, resulting in increasing training costs and spoilage of eventual predictions. In this work, we propose a unifying energy-based theory and framework called Efficient Energy-Based Expansion and Fusion (3EF) 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, 3EF extends the exemplar-set to a more challenging setting, where exemplars are randomly selected and imbalanced, where 3EF maintains its performance when prior methods fail dramatically. Extensive experiments on three widely used benchmarks: CIFAR-100, ImageNet-100, and ImageNet-1000 demonstrate that 3EF achieves state-of-the-art performance in both the ordinary and challenging CIL settings.

Out-of-distribution Representation Learning for Time Series Classification

- Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, Xing Xie
- abstract@open-review(Poster): Time series classification is an important problem in 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 inside a time series that need to be characterized. To this end, we propose DIVERSIFY for out-of-distribution (OOD) representation learning. DIVERSIFY is an end-to-end approach that takes an iterative process: it first obtains the 'worst-case' distribution scenario via adversarial training, then matches the distributions of the latent distributions. We further present some theoretical insights. 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.

Schema Inference for Interpretable Image Classification

- Haofei Zhang, Xiaokang Liu, Mengqi Xue, Kaixuan Chen, Jie Song, Mingli Song
- abstract@open-review(Poster): 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 concept of an image with the pre-computed scene impression, by analogy with the 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 and model will be made publicly available.

Your Contrastive Learning Is Secretly Doing Stochastic Neighbor Embedding

- Tianyang Hu, Zhili LIU, Fengwei Zhou, Wenjia Wang, Weiran Huang
- abstract@open-review(Poster): 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 preserving pairwise distances. In 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.

Harnessing Mixed Offline Reinforcement Learning Datasets via Trajectory Weighting

- Zhang-Wei Hong, Remi Tachet des Combes, Pulkit Agrawal, Romain Laroche
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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%).

Spiking Convolutional Neural Networks for Text Classification

- Changze Lv, Jianhan Xu, Xiaoqing Zheng
- abstract@open-review(Poster): Spiking neural networks (SNNs) offer a promising pathway to implement deep neural networks (DNNs) in a more energy-efficient manner since their neurons are sparsely 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.

Distributionally Robust Recourse Action

- Duy Nguyen, Ngoc Bui, Viet Anh Nguyen
- abstract@open-review(Poster): 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.

Write and Paint: Generative Vision-Language Models are Unified Modal Learners

- Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, Jiawei Wang
- abstract@open-review(Poster): 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 symmetrical 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 dataset 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 will be released in the final version.

Progressive Voronoi Diagram Subdivision Enables Accurate Data-free Class-Incremental Learning

- Chunwei Ma, Zhanghexuan Ji, Ziyun Huang, Yan Shen, Mingchen Gao, Jinhui Xu
- abstract@open-review(Poster): Data-free Class-incremental Learning (CIL) is a challenging problem because rehearsing data from previous phases is strictly prohibited, causing catastrophic forgetting 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
- abstract@open-review(Poster): Many recent works on understanding deep learning try to quantify how much individual data instances influence the optimization and generalization of a model, either by analyzing the behavior of the model during training or by measuring the performance gap of the model when the instance is removed from the dataset. Such approaches 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.

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
- abstract@open-review(Poster): 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 propose a novel editing algorithm to efficiently generate positive amino acid mutations that improve thermostability. Codes and datasets will be publicly released.

RPM: Generalizable Behaviors for Multi-Agent Reinforcement Learning

- Wei Qiu, Xiao Ma, Bo An, Svetlana Obraztsova, Shuicheng YAN, Zhongwen Xu
- abstract@open-review(Poster): Despite the recent advancement in multi-agent reinforcement learning (MARL), the MARL agents easily overfit the training environment and perform poorly in the 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 problem with Markov Games and propose a simple yet effective method, ranked policy memory (RPM), to collect diverse multi-agent trajectories for training MARL policies with good generalizability. The main idea of RPM is to maintain a look-up memory of policies. In particular, we try to acquire various levels of behaviors by saving policies via ranking the training episode return, i.e., the episode return of agents in the training environment; when an episode starts, the learning agent can then choose a policy from the RPM as the behavior policy. This innovative novel self-play training framework leverages agents' past policies and guarantees the diversity of multi-agent interaction in the training data. We implement RPM on top of MARL algorithms and conduct extensive experiments on Melting Pot. It has been demonstrated that RPM enables MARL agents to interact with unseen agents in multi-agent generalization evaluation scenarios and complete given tasks, and it significantly boosts the performance up to 402% 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): Deep neural networks (DNNs) are vulnerable to backdoor attacks, where adversaries 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.

On the Perils of Cascading Robust Classifiers

- Ravi Mangal, Zifan Wang, Chi Zhang, Klas Leino, Corina Pasareanu, Matt Fredrikson
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): Pretrained language models (PLMs) 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 PLMs 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 correct 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 PLM sizes.

Adaptive Robust Evidential Optimization For Open Set Detection from Imbalanced Data

- Hitesh Sapkota, Qi Yu
- abstract@open-review(Poster): Open set detection (OSD) aims at identifying data samples of an unknown class ($i.e.$, open set) from those of known classes ($i.e.$, close set) based on a model trained from close set samples. However, a close set may involve a highly imbalanced class distribution. Accurately differentiating open set samples and those from a minority class in the close 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 close 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 close and open sets, respectively, and the detection results outperform the competitive baselines.

Dual Diffusion Implicit Bridges for Image-to-Image Translation

- Xuan Su, Jiaming Song, Chenlin Meng, Stefano Ermon
- abstract@open-review(Poster): Common image-to-image translation methods rely on joint training over data from both source and target domains. This prevents the training process from preserving privacy of domain data (e.g., in a federated setting), and often means that a new model has to be trained for a new pair of domains. 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 an ODE, and 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 Schrödinger Bridges, a form of entropy-regularized optimal transport, to explain the efficacy of the method. Experimentally, we apply DDIBs on both synthetic and high-resolution image datasets, to demonstrate their utility in a wide variety of translation tasks as well as their connections to existing optimal transport methods.

Average Sensitivity of Decision Tree Learning

- Satoshi Hara, Yuichi Yoshida
- abstract@open-review(Poster): 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. A drastic change in the learned tree structure owing to such data perturbation is unfavorable in practice. For example, in data mining, a change in the tree implies that the extracted knowledge can be unstable, 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 nearly equal to the optimal decision tree. The experimental results on real-world datasets demonstrate that the proposed algorithm achieves a low average sensitivity with an insignificant decrease in accuracy.

Stable Target Field for Reduced Variance Score Estimation

- Yilun Xu, Shangyuan Tong, Tommi S. Jaakkola
- abstract@open-review(Poster): Score-based generative models (SGMs) generate samples by reversing a fixed forward diffusion process. Despite impressive empirical results, we observe that the training process leads to unstable outcomes, especially when the reverse-time solvers adopt a large step size. The performance of converged models varies significantly with different random seeds, and they produce noticeable artifacts in generated samples. We suggest that the source of such instability lies in the handling of intermediate noise-variance scales, where multiple modes in the data affect the direction of reverse paths. Thus, the score-matching objective has a large sample variance in this regime, explaining lesser quality score estimates. We propose to remedy the problem by incorporating a reference batch for minibatch updates where the reference batch is used to calculate weighted conditional scores as the 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 of state-of-the-art SGMs across datasets with both general ODE and SDE solvers. In particular, our method improves and stabilizes the final performance of SGMs, as well as speeding up the training process.

Countinuous pseudo-labeling from the start

- Dan Berrebbi, Ronan Collobert, Samy Bengio, Navdeep Jaitly, Tatiana Likhomanenko
- abstract@open-review(Poster): 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.

GNNDelete: A General Unlearning Strategy for Graph Neural Networks

- Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, Marinka Zitnik
- abstract@open-review(Poster): We consider the problem of graph unlearning, wherein graph neural network (GNN) model is trained to specified accuracy and then deployed while a sequence of requests arrives to delete graph elements (nodes, edges) from the model. As GNN models are used in real-world implementation, this problem is increasingly vital to address—for example, when a user seeks to hide their connections with others in a social graph or when relationships in a knowledge graph become irrelevant or are not true anymore. To unlearn information from a trained GNN, its influence on both GNN model weights as well as on representations of neighbors in the graph must be deleted from the model. However, existing methods using retraining and weight modification either degrade model weights shared across all nodes or are ineffective because of strong dependency of deleted edges on their local graph neighborhood. Realizing these pitfalls, we formalize the required properties for graph unlearning in the form of Deleted Edge Consistency and Neighborhood Influence and develop GNNDELETE, a model-agnostic layer-wise operator that optimize both properties for unlearning tasks. GNNDELETE updates latent representations to delete nodes and edges from the model while keeping the rest of the learned knowledge intact. Experiments on six real-world and two knowledge graphs show that GNNDELETE outperforms existing graph unlearning models by up to 36.9% in AUC on link prediction tasks and 22.5% in AUC on distinguishing deleted edges from nondeleted edges. GNNDELETE efficient—e.g., it takes 12.3x less time and 9.3x less space than retraining from scratch on WordNet18. Our code is available here.

Meta-Learning in Games

- Keegan Harris, Ioannis Anagnostides, Gabriele Farina, Mikhail Khodak, Steven Wu, Tuomas Sandholm
- abstract@open-review(Poster): 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.

DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking

- Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, Tommi S. Jaakkola
- abstract@open-review(Poster): 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<2Å) on PDBBind, significantly outperforming the previous state-of-the-art of traditional docking (23%) and deep learning (20%) methods. Moreover, DiffDock has fast inference times and provides confidence estimates with high selective accuracy.

Constructive TT-representation of the tensors given as index interaction functions with applications

- Gleb Ryzhakov, Ivan Oseledets
- abstract@open-review(Poster): 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.

Sparse tree-based Initialization for Neural Networks

- Patrick Lutz, Ludovic Arnould, Claire Boyer, Erwan Scornet
- abstract@open-review(Poster): 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 stochastic gradient strategies. Numerical experiments on several tabular data sets show that this new, simple and easy-to-use method is a solid concurrent, both in terms of generalization capacity and computation time, to default MLP initialization and even to existing complex deep learning solutions. In fact, this wise MLP initialization raises the resulting NN methods to the level of a valid competitor to gradient boosting when dealing with tabular data. Besides, such initializations are able to preserve the sparsity of weights introduced in the first layers of the network through training. This fact suggests that this new initializer operates an implicit regularization during the NN training, and emphasizes that the first layers act as a sparse feature extractor (as for convolutional layers in CNN).

VoGE: A Differentiable Volume Renderer using Gaussian Ellipsoids for Analysis-by-Synthesis

- Angtian Wang, Peng Wang, Jian Sun, Adam Kortylewski, Alan Yuille
- abstract@open-review(Poster): 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 will be publicly available.

On Emergence of Activation Sparsity in Trained 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

- abstract@open-review(Poster): This paper reveals a curious observation that modern large-scale machine learning models 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 value of k brings a collection of desired but missing properties for Transformers, namely less sensitivity to noisy training data, more robustness to input corruptions, and better calibration for their prediction confidence.

FoSR: First-order spectral rewiring for addressing oversquashing in GNNs

- Kedar Karhadkar, Pradeep Kr. Banerjee, Guido Montufar
- abstract@open-review(Poster): 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.

Generative Modeling Helps Weak Supervision (and Vice Versa)

- Benedikt Boecking, Nicholas Roberts, Willie Neiswanger, Stefano Ermon, Frederic Sala, Artur Dubrawski
- abstract@open-review(Poster): 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.

Provable Memorization Capacity of Transformers

- Junghwan Kim, Michelle Kim, Barzan Mozafari
- abstract@open-review(Poster): 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 task. To verify these theoretical findings, we conduct experiments analyzing the memorization capacity of Transformers in the natural language domain.

Bridge the Inference Gaps of Neural Processes via Expectation Maximization

- Qi Wang, Marco Federici, Herke van Hoof
- abstract@open-review(Poster): 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.

Masked Vision and Language Modeling for Multi-modal Representation Learning

- Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, Stefano Soatto
- abstract@open-review(Poster): 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.

Agent-based Graph Neural Networks

- Karolis Martinkus, Pál András Papp, Benedikt Schesch, Roger Wattenhofer
- abstract@open-review(Poster): 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 3-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.

On the Performance of Temporal Difference Learning With Neural Networks

- HAOXING TIAN, Ioannis Paschalidis, Alex Olshevsky
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 this 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 two 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 framework, we are able to identify a novel chemical space coverage measure, #Circles, 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.

Understanding new tasks through the lens of training data via exponential tilting

- Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, Yuekai Sun
- abstract@open-review(Poster): 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.

Calibrating Sequence likelihood Improves Conditional Language Generation

- Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, Peter J Liu
- abstract@open-review(Poster): 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.

Learning differentiable solvers for systems with hard constraints

- Geoffrey Négiar, Michael W. Mahoney, Aditi Krishnapriyan
- abstract@open-review(Poster): We introduce a practical method to enforce linear 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
- abstract@open-review(Poster): 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 clients' 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.

SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models

- Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, Animesh Garg

- abstract@open-review(Poster): 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
- abstract@open-review(Poster): While reinforcement learning (RL) methods that learn an internal model of the environment 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 such sample efficient methods typically are computationally demanding, our method attains the performance of SAC in about 50% less wall-clock time.

Deep Generative Symbolic Regression

- Samuel Holt, Zhaozhi Qian, Mihaela van der Schaar
- abstract@open-review(Poster): 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 inference 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
- abstract@open-review(Poster): 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.

Predictor-corrector algorithms for stochastic optimization under gradual distribution shift

- Subha Maity, Debarghya Mukherjee, Moulinath Banerjee, Yuekai Sun
- abstract@open-review(Poster): 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 the 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 Understanding

- Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, Mu Li
- abstract@open-review(Poster): 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, full 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.

Impossibly Good Experts and How to Follow Them

- Aaron Walsman, Muru Zhang, Sanjiban Choudhury, Dieter Fox, Ali Farhadi
- abstract@open-review(Poster): 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 environments.

Distributionally Robust Post-hoc Classifiers under Prior Shifts

- Jiaheng Wei, Harikrishna Narasimhan, Ehsan Amid, Wen-Sheng Chu, Yang Liu, Abhishek Kumar
- abstract@open-review(Poster): The generalization ability of machine learning models degrades significantly when the test distribution shifts away from the training distribution. We investigate 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.

Transformer Meets Boundary Value Inverse Problems

- Ruchi Guo, Shuhao Cao, Long Chen
- abstract@open-review(Poster): A Transformer-based deep direct sampling method is proposed for solving a class of 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 but critical question: whether and how one can benefit from the theoretical structure of a mathematical problem to develop task-oriented and structure-conforming deep neural network? Specifically, inspired by direct sampling methods for inverse problems, the 1D boundary data are preprocessed by a partial differential equation-based feature map to yield 2D harmonic extensions in different frequencies as different input channels. Then, by introducing learnable non-local kernel, the approximation of direct sampling is recast to a modified attention mechanism. The proposed method is then applied to electrical impedance tomography, a well-known severely ill-posed nonlinear inverse problem. The new method achieves superior accuracy over its predecessors and contemporary operator learners, as well as shows robustness with respect to noise. This research shall strengthen the insights that the attention mechanism, despite being invented for natural language processing tasks, 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.

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
- abstract@open-review(Poster): 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 dataset 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 conflicts. To alleviate such conflicts, 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, respectively, making the classification conflict-robust and the feature embedding compact. Our method outperforms state-of-the-art unsupervised and supervised image retrieval approaches on multiple benchmarks with substantial improvement under different dimension constraints. Pre-processed data, training code, and pre-trained models will be released to reproduce our results.

Diffusion Probabilistic Fields

- Peiye Zhuang, Samira Abnar, Jiatao Gu, Alex Schwing, Joshua M. Susskind, Miguel Ángel Bautista
- abstract@open-review(Poster): 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 an 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 and end-to-end learning algorithm that side-steps the requirement of representing fields with latent vectors as in previous approaches. 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.

Beyond calibration: estimating the grouping loss of modern neural networks

- Alexandre Perez-Lebel, Marine Le Morvan, Gael Varoquaux
- abstract@open-review(Poster): Good decision making requires machine-learning models to provide trustworthy confidence scores. To this end, recent work has focused on miscalibration, i.e, the over or under confidence of model scores. Yet, contrary to widespread belief, calibration is not enough: even a classifier with the best possible accuracy and perfect calibration can have confidence scores 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 use it to study modern neural network architectures in vision and NLP. We find that the grouping loss varies markedly across architectures, and that it is a key model-comparison factor across the most accurate, calibrated, models. We also show that distribution shifts lead to high grouping loss.

Hybrid RL: Using both offline and online data can make RL efficient

- Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, Wen Sun
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

GAIN: On the Generalization of Instructional Action Understanding

- Junlong Li, Guangyi Chen, Yansong Tang, Jinan Bao, Kun Zhang, Jie Zhou, Jiwen Lu
- abstract@open-review(Poster): 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.

ManyDG: Many-domain Generalization for Healthcare Applications

- Chaoqi Yang, M Brandon Westover, Jimeng Sun
- abstract@open-review(Poster): 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 separate 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Robust Attention for Contextual Biased Visual Recognition

- Ruyang Liu, Jingjia Huang, Ge Li, Thomas H. Li
- abstract@open-review(Poster): 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 protects 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 multi-label classification on MS-COCO and PASCAL-VOC. The codes will be publicly available.

Multi-Objective Reinforcement Learning: Convexity, Stationarity and Pareto Optimality

- Haoye Lu, Daniel Herman, Yaoliang Yu
- abstract@open-review(Poster): 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.

Fooling SHAP with Stealthily Biased Sampling

- gabriel laberge, Ulrich Aïvodji, Satoshi Hara, Mario Marchand, Foutse Khomh
- abstract@open-review(Poster): 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. 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 model 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
- abstract@open-review(Poster): 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.

Trading Information between Latents in Hierarchical Variational Autoencoders

- Tim Z. Xiao, Robert Bamler
- abstract@open-review(Poster): 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.

ISAAC Newton: Input-based Approximate Curvature for Newton's Method

- Felix Petersen, Tobias Sutter, Christian Borgelt, Dongsung Huh, Hilde Kuehne, Yuekai Sun, Oliver Deussen
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 super- vised 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 rep- resentations. 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
- abstract@open-review(Poster): 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.

How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection?

- Yifei Ming, Yiyou Sun, Ousmane Dia, Yixuan Li
- abstract@open-review(Poster): 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.

AnyDA: Anytime Domain Adaptation

- Omprakash Chakraborty, Aadarsh Sahoo, Rameswar Panda, Abir Das
- abstract@open-review(Poster): 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 limited to testing under a specific target setting. This can limit their impact for many real-world applications that present different resource constraints. 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. Extensive experiments on several diverse benchmark datasets well demonstrate the superiority of our proposed approach over state-of-the-art methods.

Improving Deep Regression with Ordinal Entropy

- Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, Angela Yao
- abstract@open-review(Poster): In computer vision, it is often observed that formulating regression problems as a classification task often 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.

Unified Discrete Diffusion for Simultaneous Vision-Language Generation

- Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, Ponnuthurai N. Suganthan
- abstract@open-review(Poster): 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.

Iterative Patch Selection for High-Resolution Image Recognition

- Benjamin Bergner, Christoph Lippert, Aravindh Mahendran
- abstract@open-review(Poster): 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.

Fuzzy Alignments in Directed Acyclic Graph for Non-Autoregressive Machine Translation

- Zhengrui Ma, Chenze Shao, Shangtong Gui, Min Zhang, Yang Feng
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

3D Segmenter: 3D Transformer based Semantic Segmentation via 2D Panoramic Distillation

- ZHENNAN WU, YANG LI, Yifei Huang, Lin Gu, Tatsuya Harada, Hiroyuki Sato
- abstract@open-review(Poster): 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 decoder. 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.

Clifford Neural Layers for PDE Modeling

- Johannes Brandstetter, Rianne van den Berg, Max Welling, Jayesh K Gupta
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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-based 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%.

TabCaps: A Capsule Neural Network for Tabular Data Classification with BoW Routing

- Jintai Chen, KuanLun Liao, Yanwen Fang, Danny Chen, Jian Wu
- abstract@open-review(Poster): The instances in a table are represented by a collection of heterogeneous tabular features. Previous work often made predictions for such instances in a paradigm that processed tabular features as operating units, which requires to well cope with the heterogeneity. In this paper, we propose to encapsulate all tabular features of an instance into vectorial features and process them collectively rather than have to deal with individual ones, which directly captures the representations at the instance level and benefits robust performances. Specifically, we adopt "capsules" to organize tabular features of the instance into vectorial features, and devise a novel capsule neural network called TabCaps to process the vectorial features for classification. In TabCaps, a tabular instance is respectively 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 instance and is transformed into senior capsule layer under the guidance of a novel 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.

An Exact Poly-Time Membership-Queries Algorithm for Extracting a Three-Layer ReLU Network

- Amit Daniely, Elad Granot
- abstract@open-review(Poster): 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.

Towards Understanding and Mitigating Dimensional Collapse in Heterogeneous Federated Learning

- Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, Song Bai
- abstract@open-review(Poster): 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/Yujun-Shi/FedCLS.

Evidential Uncertainty and Diversity Guided Active Learning for Scene Graph Generation

- Shuzhou Sun, Shuaifeng Zhi, Janne Heikkilä, Li Liu
- abstract@open-review(Poster): 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.

Anisotropic Message Passing: Graph Neural Networks with Directional and Long-Range Interactions

- Moritz Thürlemann, Sereina Riniker
- abstract@open-review(Poster): 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 seamless 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.

SYNC: SAFETY-AWARE NEURAL CONTROL FOR STABILIZING STOCHASTIC DELAY-DIFFERENTIAL EQUATIONS

- Jingdong Zhang, Qunxi Zhu, Wei Yang, Wei Lin
- abstract@open-review(Poster): Stabilization of the systems described by stochastic delay-differential equations is a challenging task in 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. The two learned control policies, the neural deterministic controller (NDC) and the neural stochastic controller (NSC), work effectively because the learning procedures use, 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 NDC and NSC 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 can only evolve in some specific safety set. Such successful stabilization based on neural networks restricted in safety set is attributed to our further developed theory for safety verification of SDDEs using the stochastic control barrier function, and we name it 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.

Differentiable Mathematical Programming for Object-Centric Representation Learning

- Adeel Pervez, Phillip Lippe, Efstratios Gavves
- abstract@open-review(Poster): 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 results show that our approach is scalable and outperforms existing methods on object discovery tasks with textured scenes and objects.

Scalable Subset Sampling with Neural Conditional Poisson Networks

- Adeel Pervez, Phillip Lippe, Efstratios Gavves
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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-tailed, 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.

Learning to Estimate Single-View Volumetric Flow Motions without 3D Supervision

- Erik Franz, Barbara Solenthaler, Nils Thuerey
- abstract@open-review(Poster): 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.

Towards the Out-of-Distribution Generalization of Contrastive Self-Supervised Learning

- Xuyang Zhao, Tianqi Du, Yisen Wang, Jun Yao, Weiran Huang
- abstract@open-review(Poster): Self-supervised learning attracts much attention recently, since it does not require labeled data for training contrasted to supervised learning. Empirical studies also observe that it has better transfer ability than supervised learning. However, the theoretical study of the out-of-distribution (OOD) generalization ability of self-supervised learning is still limited. In this paper, by focusing on the data augmentation used in SSL, we establish a theoretical framework for the OOD performance of contrastive-based self-supervised learning. Although some recent work claims that contrastive learning learns more robust representations than supervised learning, our results suggest that this superiority mainly comes from the data augmentation used, i.e., more data are fed to the model. In the face of more challenging OOD scenarios, the standard contrastive learning still suffers from the same generalization problem as empirical risk minimization (ERM). Based on our theoretical results, we propose an augmentation-robust contrastive learning approach, named as ArCL, which significantly improves the OOD performance of contrastive learning in several datasets.

Temperature Schedules for self-supervised contrastive methods on long-tail data

- Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, Christian Rupprecht
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

ImaginaryNet: Learning Object Detectors without Real Images and Annotations

- Minheng Ni, Zitong Huang, Kailai Feng, Wangmeng Zuo
- abstract@open-review(Poster): Humans can easily detect a known concept without the demand of training in reality. Equipping this ability to deep learning may allow the neural network to learn complex vision models, e.g., object detection, without collecting and annotating real images. In this paper, we define a novel paradigm as Imaginary-Supervised Object Detection (ISOD), where no real images and manual annotations are used for training object detectors. To resolve this challenge, we propose ImaginaryNet, a framework to learn object detectors by combining pretrained language model as well as text-to-image synthesis models. In particular, photo-realistic images can be generated by the text-to-image model, and class labels can be obtained by the text generated by the language model. Then, weakly supervised object detection is leveraged to learn the detector without real images and manual annotations. 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 70% 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 real images and manual annotations.

Contextual bandits with concave rewards, and an application to fair ranking

- Virginie Do, Elvis Dohmatob, Matteo Pirotta, Alessandro Lazaric, Nicolas Usunier
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

The Continuous CNN: from Task-Specific to Unified CNN Architecture

- David M Knigge, David W. Romero, Albert Gu, Efstratios Gavves, Erik J Bekkers, Jakub Mikolaj Tomczak, Mark Hoogendoorn, Jan-jakob Sonke
- abstract@open-review(Poster): Performant Convolutional Neural Network (CNN) architectures must be tailored to specific tasks in order to incorporate considerations such as input length, resolution, and dimentionality of the data. To overcome the need for such problem-specific CNN architectures, and the fragmentation they represent to the field, we introduce the \textit{Continuous Convolutional Neural Network} (CCNN): a single CNN architecture that can be used for tasks on data of arbitrary resolution, dimensionality and length without structural changes. The key component of the CCNN is its \textit{continuous convolutional kernel} which models long-range dependencies at every layer and removes the need for downsampling and task-dependent depths used in current CNN architectures. We demonstrate the generality of our CCNN by deploying the \emph{same architecture} to tasks on sequential ($1{\rm D}$), visual ($2{\rm D}$), and point-cloud ($3{\rm D}$) data. Experiments show that the CCNN matches and often outperforms the current state-of-the-art across the tasks considered.

Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training

- Simone Zini, Alex Gomez-Villa, Marco Buzzelli, Bartłomiej Twardowski, Andrew D. Bagdanov, Joost van de weijer
- abstract@open-review(Poster): 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.

GAMR: A Guided Attention Model for (visual) Reasoning

- Mohit Vaishnav, Thomas Serre
- abstract@open-review(Poster): Humans continue to outperform modern AI systems in their ability to flexibly parse and understand complex visual scenes. Here, we present a novel transformer-based 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
- abstract@open-review(Poster): 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.

Approximate Nearest Neighbor Search through Modern Error-Correcting Codes

- Noam Touitou, Nissim Halabi
- abstract@open-review(Poster): 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.

When to Make and Break Commitments?

- Alihan Hüyük, Zhaozhi Qian, Mihaela van der Schaar
- abstract@open-review(Poster): 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.

DENSE RGB SLAM WITH NEURAL IMPLICIT MAPS

- Heng Li, Xiaodong Gu, Weihao Yuan, luwei yang, Zilong Dong, Ping Tan
- abstract@open-review(Poster): 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 benchmark datasets and compare 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. Our source code will be publicly available.

Monocular Scene Reconstruction with 3D SDF Transformers

- Weihao Yuan, Xiaodong Gu, Heng Li, Zilong Dong, Siyu Zhu
- abstract@open-review(Poster): Monocular scene reconstruction from posed images is challenging due to the complexity of a large environment. Recent volumetric methods learn to directly predict 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

D4AM: A General Denoising Framework for Downstream Acoustic Models

- Chi-Chang Lee, Yu Tsao, Hsin-Min Wang, Chu-Song Chen
- abstract@open-review(Poster): The performance of acoustic models degrades notably in noisy environments. Speech enhancement (SE) can be used as a front-end strategy to serve automatic speech recognition (ASR) systems. However, the training objectives of existing SE approaches do not consider the generalization ability to unseen ASR systems. In this study, we propose a general denoising framework for various downstream acoustic models, called D4AM. Our framework fine-tunes the SE model with the backward gradient according to a specific acoustic model and the corresponding classification objective. At the same time, our method aims to take 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 instead of going through a grid search process with additional training costs. The adjustment scheme consists of two parts: gradient calibration and regression objective weighting. Experimental results show that D4AM can consistently and effectively provide improvements to various unseen acoustic models and outperforms other combination setups. To the best of 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.

Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning

- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, Tuo Zhao
- abstract@open-review(Poster): 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 MARVEL, which adaptively allocates the parameter budget among weight matrices according to their importance score. In particular, MARVEL 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 MARVEL. Results demonstrate that MARVEL manifests notable improvement over baselines, especially in the low budget settings. Our code will be publicly available.

Generalize Learned Heuristics to Solve Large-scale Vehicle Routing Problems in Real-time

- Qingchun Hou, Jingwei Yang, Yiqiang Su, Xiaoqing Wang, Yuming Deng
- abstract@open-review(Poster): Large-scale Vehicle Routing Problems (VRPs) are widely used in logistics, transportation, supply chain, and robotic system. Recently, data-driven VRP heuristics are proposed to generate real-time VRP solutions with up to 100 nodes. However, current heuristics for large-scale VRPs still face three challenges: 1) Hard to generalize the heuristics learned on small-scale VRPs to large-scale VRPs in zero-shot way; 2) Hard to generate real-time solutions for large-scale VRPs; 3) Hard to embed global constraints in 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.

Towards the Generalization of Contrastive Self-Supervised Learning

- Weiran Huang, Mingyang Yi, Xuyang Zhao, Zihao Jiang
- abstract@open-review(Poster): 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 performance. 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 can be optimized by contrastive algorithms, while the third one is priorly determined by pre-defined data augmentation. With the above theoretical findings, we then study two canonical contrastive losses, InfoNCE and cross-correlation, to see how they satisfy the first two factors. Furthermore, we conduct various experiments to study the third factor, and observe that the downstream performance is highly correlated to the concentration of augmented data.

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
- abstract@open-review(Poster): 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 infeasible 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Advancing Radiograph Representation Learning with Masked Record Modeling

- Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, Yizhou Yu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Re-parameterizing Your Optimizers rather than Architectures

- Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Kaiqi Huang, Jungong Han, Guiguang Ding
- abstract@open-review(Poster): 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. We will make the code and models publicly available.

Protein Representation Learning via Knowledge Enhanced Primary Structure Reasoning

- Hong-Yu Zhou, Yunxiang Fu, Zhicheng Zhang, Bian Cheng, Yizhou Yu
- abstract@open-review(Poster): 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 proteins and associated knowledge terms as the knowledge encoding objective. However, it fails to consider the semantic gap between protein sequences and natural language, and the resulting feature misalignment may adversely affect representation learning. To mitigate this, we propose Knowledge-exploited Auto-encoder for Proteins (KeAP), which performs implicit knowledge encoding by learning to exploit knowledge for protein primary structure reasoning. In practice, the protein representation iteratively queries the associated knowledge terms to extract and integrate helpful information for restoring missing amino acids via attention, avoiding a direct comparison between the two modalities. 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 encoding in protein representation learning.

Provable Unsupervised Data Sharing for Offline Reinforcement Learning

- Hao Hu, Yiqin Yang, Qianchuan Zhao, Chongjie Zhang
- abstract@open-review(Poster): Self-supervised methods play a vital role in fueling the progress of deep learning using supervision from the data itself, obviating the need for expensive annotations. The same merit applies to offline reinforcement learning (RL), which conducts RL in a supervised manner, but it is unclear how to utilize such unlabeled data to improve offline RL in a principled way. In this paper, we examine the theoretical benefit of unlabeled data in the context of linear MDPs and propose a novel and Provable Data Sharing algorithm, which we refer to as PDS, to utilize such unlabeled data for offline RL. PDS utilizes additional penalties upon the reward function learned from labeled data to avoid potential overestimation of the reward. We show that such a penalty is crucial to keep the algorithm conservative, and PDS achieves a provable benefit from unlabeled data under mild conditions. We conduct extensive experiments on various offline RL tasks and show that PDS can significantly improve offline RL algorithms with unlabeled data.

Modeling Sequential Sentence Relation to Improve Cross-lingual Dense Retrieval

- Shunyu Zhang, Yaobo Liang, MING GONG, Daxin Jiang, Nan Duan
- abstract@open-review(Poster): 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. Code and model will be available.

DepthFL : Depthwise Federated Learning for Heterogeneous Clients

- Minjae Kim, Sangyoon Yu, Suhyun Kim, Soo-Mook Moon
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 denoising 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.

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
- abstract@open-review(Poster): 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-team 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 anonymized-gobigger.github.io.

Masked Unsupervised Self-training for Label-free Image Classification

- Junnan Li, Silvio Savarese, Steven Hoi
- abstract@open-review(Poster): 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 submitted in the supplementary material.

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, Harry Shum
- abstract@open-review(Poster): 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.

Revisiting Graph Adversarial Attack and Defense From a Data Distribution Perspective

- Kuan Li, Yang Liu, Xiang Ao, Qing He
- abstract@open-review(Poster): 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.

Provable Sim-to-real Transfer in Continuous Domain with Partial Observations

- Jiachen Hu, Han Zhong, Chi Jin, Liwei Wang
- abstract@open-review(Poster): 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.

Globally Optimal Training of Neural Networks with Threshold Activation Functions

- Tolga Ergen, Halil Ibrahim Gulluk, Jonathan Lacotte, Mert Pilanci
- abstract@open-review(Poster): 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 an alternative simplified convex optimization formulation when the set of hyperplane arrangements for the data matrix is complete, i.e., 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, Qi Liu, Shuxin Zheng, Yaosen Min
- abstract@open-review(Poster): 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.

Towards Robustness Certification Against Universal Perturbations

- Yi Zeng, Zhouxing Shi, Ming Jin, Feiyang Kang, Lingjuan Lyu, Cho-Jui Hsieh, Ruoxi Jia
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Basic Binary Convolution Unit for Binarized Image Restoration Network

- Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, Luc Van Gool
- abstract@open-review(Poster): 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.

Multimodal Federated Learning via Contrastive Representation Ensemble

- Qiying Yu, Yimu Wang, Ke Xu, Yang Liu, Jingjing Liu
- abstract@open-review(Poster): 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 superiority 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
- abstract@open-review(Poster): 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.

Can CNNs Be More Robust Than Transformers?

- Zeyu Wang, Yutong Bai, Yuyin Zhou, Cihang Xie
- abstract@open-review(Poster): 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 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 is 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.

Risk-Aware Reinforcement Learning with Coherent Risk Measures and Non-linear Function Approximation

- Thanh Lam, Arun Verma, Bryan Kian Hsiang Low, Patrick Jaillet
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

On the Saturation Effect of Kernel Ridge Regression

- Yicheng Li, Haobo Zhang, Qian Lin
- abstract@open-review(Poster): 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.

Protein Representation Learning by Geometric Structure Pretraining

- Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, Jian Tang
- abstract@open-review(Poster): Learning effective representations of proteins 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 structure encoder by leveraging multiview contrastive learning and compare against pretraining with various 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 data. All codes and models will be published upon acceptance.

Trainable Weight Averaging: Efficient Training by Optimizing Historical Solutions

- Tao Li, Zhehao Huang, Qinghua Tao, Yingwen Wu, Xiaolin Huang
- abstract@open-review(Poster): 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.

Deep Declarative Dynamic Time Warping for End-to-End Learning of Alignment Paths

- Ming Xu, Sourav Garg, Michael Milford, Stephen Gould
- abstract@open-review(Poster): This paper addresses end-to-end learnable 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 continuous relaxation to the min operator found in the recursive steps used to solve the DTW problem. We instead propose a DTW layer based around deep declarative networks. By formulating the DTW problem as a continuous, inequality constrained optimisation problem, we can compute exact gradients for the solution of the optimal alignment (with respect to the underlying time series) using implicit differentiation. Our formulation yields a major improvement over existing approaches; our DTW layer outputs the entire warping path between two time series as opposed to only the DTW discrepancy value. This enables the specification of downstream loss functions on the alignment path itself, useful for instance, to improve the precision of predicted alignments when ground-truth alignments are available. We evaluate our method 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.

Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning

- Ting Chen, Ruixiang ZHANG, Geoffrey Hinton
- abstract@open-review(Poster): We present Bit Diffusion: a simple and generic approach for generating discrete data with continuous 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
- abstract@open-review(Poster): 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/η (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 give 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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).

Neural Radiance Field Codebooks

- Matthew Wallingford, Aditya Kusupati, Alex Fang, Vivek Ramanujan, Aniruddha Kembhavi, Roozbeh Mottaghi, Ali Farhadi
- abstract@open-review(Poster): 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.

Scalable Estimation of Nonparametric Markov Networks with Mixed-Type Data

- Yujia Zheng, Ignavier Ng, Yewen Fan, Kun Zhang
- abstract@open-review(Poster): 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 the graph, and most of them can only handle variables of a single data type (continuous or discrete). In this work, we generalize the characterization of the conditional independence structure to handle general distributions for all data types (i.e., continuous, discrete, and mixed-type) with 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 experimentally and demonstrate the scalability of the approach--it produces the estimated Markov network over up to 5000 nodes within one hour on CPUs. We further discuss the implication of the proposed approach in causal discovery.

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
- abstract@open-review(Poster): 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.

Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation

- Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, Yan Yan
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 segmentation 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.

Rethinking Graph Lottery Tickets: Graph Sparsity Matters

- Bo Hui, Da Yan, Xiaolong Ma, Wei-Shinn Ku
- abstract@open-review(Poster): 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.

Private Federated Learning Without a Trusted Server: Optimal Algorithms for Convex Losses

- Andrew Lowy, Meisam Razaviyayn
- abstract@open-review(Poster): 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 bounds 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.

Cheap Talk Discovery and Utilization in Multi-Agent Reinforcement Learning

- Yat Long Lo, Christian Schroeder de Witt, Samuel Sokota, Jakob Nicolaus Foerster, Shimon Whiteson
- abstract@open-review(Poster): 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.

Reversible Column Networks

- Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, Xiangyu Zhang
- abstract@open-review(Poster): We propose a new neural network design paradigm Reversible Column Networks (RevCols). The main body of RevCols is composed of multiple copies of subnetworks, named columns respectively, between which multi-level reversible connections are employed. Such architectural scheme attributes RevCols very different behavior from conventional networks: during forward propagation, features in RevCols 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 RevCols 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, 61.2% APbox and 53.6% APmask on COCO detection test-dev set, 57.1% mIoU on ADE20k segmentation. To our knowledge, it is the best COCO detection result among pure CNN models without extra detection data. Moreover, as a general macro architecture fashion, RevCols can also be introduced into transformers or other neural networks, which is demonstrated to improve the performances in both computer vision and NLP tasks.

Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Experts

- Zhitong Gao, Yucong Chen, Chuyu Zhang, Xuming He
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 state 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 proposed 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 is strong, as it can both induce unsafe behaviors and make the attack stealthy by maintaining the reward. We further propose a more effective adversarial training framework for safe RL and evaluate it via comprehensive experiments (video demos are available at: https://sites.google.com/view/robustsaferl/home). This paper provides a pioneer work to investigate the safety and robustness of RL under observational attacks for future safe RL studies.

Behind the Scenes of Gradient Descent: A Trajectory Analysis via Basis Function Decomposition

- Jianhao Ma, Lingjun Guo, Salar Fattahi
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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. In light of this, 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 performance, demonstrating that ParetoGNN achieves better task generalization via the disjoint yet complementary knowledge learned from different philosophies.

Analyzing Tree Architectures in Ensembles via Neural Tangent Kernel

- Ryuichi Kanoh, Mahito Sugiyama
- abstract@open-review(Poster): A soft tree is an actively studied variant of a decision tree that updates splitting rules using the gradient method. Although it can take various tree 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 that only the number of leaves at each depth is relevant for the tree architecture in ensemble learning with infinitely many 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.

Exploring The Role of Mean Teachers in Self-supervised Masked Auto-Encoders

- Youngwan Lee, Jeffrey Ryan Willette, Jonghee Kim, Juho Lee, Sung Ju Hwang
- abstract@open-review(Poster): 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 than MAE on downstream tasks such as ImageNet-1K classification, object detection, and instance segmentation.

Sub-Task Decomposition Enables Learning in Sequence to Sequence Tasks

- Noam Wies, Yoav Levine, Amnon Shashua
- abstract@open-review(Poster): 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.

Evaluating Long-Term Memory in 3D Mazes

- Jurgis Pašukonis, Timothy P Lillicrap, Danijar Hafner
- abstract@open-review(Poster): 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.

Proactive Multi-Camera Collaboration for 3D Human Pose Estimation

- Hai Ci, Mickel Liu, Xuehai Pan, fangwei zhong, Yizhou Wang
- abstract@open-review(Poster): For human motion capture (MoCap), particularly outdoors, the fixed-viewpoint multi-camera solutions are susceptible to dynamic occlusions and constrained in capture space. While an active camera approach aims to proactively control the camera poses to find optimal viewpoints for 3D reconstruction. This work introduces a multi-agent reinforcement learning (MARL) scheme to proactive Multi-Camera Collaboration for 3D Human Pose Estimation (MCC-HPE) in dynamic human crowds. At its core is a novel Collaborative Triangulation Contribution Reward (CTCR) that incentivizes agents according to their weighted average marginal contribution to the 3D reconstruction. CTCR improves convergence and alleviates the multi-agent credit assignment issue resulted from using 3D reconstruction accuracy as the shared reward. To better capture environment dynamics and to encourage anticipatory behaviors for occlusion avoidance, we jointly train our model with multiple world dynamics learning tasks. We evaluate our proposed method in four photo-realistic UE4 environments to ensure validity and generalizability. The empirical results show that our methods steadily outperform the fixed and active baselines in different scenarios with various numbers of cameras and humans.

Become a Proficient Player with Limited Data through Watching Pure Videos

- Weirui Ye, Yunsheng Zhang, Pieter Abbeel, Yang Gao
- abstract@open-review(Poster): 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 cycle consistency 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.

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
- abstract@open-review(Poster): We transfer motions from a target dynamic person to a source static one. An accurate matching between the source and the target human subjects improves the transferred motion quality. This matching paradigm shall be effective to capture both large and subtle motions changes between the

target person and the source person, which are challenging for existing methods with CNNs. In this paper, we propose Human MotionFormer, a hierarchical ViT framework for motion transfer between two human subjects. Our MotionFormer leverages the local and the global perceptions (via convolutions and cross-attentions) to capture varying human motions. Specifically, our MotionFormer consists of two ViT encoders to extract input features (i.e., target pose image and a source human image), and a ViT decoder with several blocks for feature matching and motion transfer. The feature matching process is conducted in both motion warping and generation branches within each decoder block, where the target pose feature is set as Query and the source person feature is set as Key and Value. Then, the cross attention maps computed based on the Query, Key and Value are utilized for feature matching. Furthermore, we introduce a convolution layer to improve the local perception after the global cross attention computations. During model training, we propose a mutual learning loss to enable the co-supervision between motion warping and generation branches for consistent motion representations, which benefit the output transferred human motions. To this end, our MotionFormer leverages the local and global perceptions and introduces the mutual learning loss to improve transferred motion results. These designs empower our method to utilize only one source image for motion transfer and get rid of model finetuning. The experimental results qualitatively and quantitatively show that our Human MotionFormer sets the new state-of-the-art performance.

Backstepping Temporal Difference Learning

- Han-Dong Lim, Donghwan Lee
- abstract@open-review(Poster): 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.

Rank Preserving Framework for Asymmetric Image Retrieval

- Hui Wu, Min Wang, Wengang Zhou, Houqiang Li
- abstract@open-review(Poster): 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 preserved well 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.

Mega: Moving Average Equipped Gated Attention

- Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, Luke Zettlemoyer
- abstract@open-review(Poster): 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 benchmarks, 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.

Parallel Deep Neural Networks Have Zero Duality Gap

- Yifei Wang, Tolga Ergen, Mert Pilanci
- abstract@open-review(Poster): 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 program for training. However, extensions of this result to deeper networks remain to be an open problem. In this paper, we particularly 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. Therefore, we prove the strong duality and existence of equivalent convex programs that enable convex and 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.

Information-Theoretic Analysis of Unsupervised Domain Adaptation

- Ziqiao Wang, Yongyi Mao
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Understanding Zero-shot Adversarial Robustness for Large-Scale Models

- Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, Carl Vondrick
- abstract@open-review(Poster): 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.

Can We Faithfully Represent Absence States to Compute Shapley Values on a DNN?

- Jie Ren, Zhanpeng Zhou, Qirui Chen, Quanshi Zhang
- abstract@open-review(Poster): Although many methods have been proposed to estimate attributions of input variables, there still exists a significant theoretical flaw in the masking-based attribution methods, i.e., it is hard to examine whether the masking method faithfully represents the absence of input variables. Specifically, for masking-based attributions, setting an input variable to the baseline value is a typical way of representing the absence of the variable. However, there are no studies investigating how to represent the absence of input variables and verify the faithfulness of baseline values. Therefore, we revisit the feature representation of a deep model in terms of causality, and propose to use causal patterns to examine whether the masking method faithfully removes information encoded in the input variable. More crucially, it is proven that the causality can be explained as the elementary rationale of the Shapley value. Furthermore, we define the optimal baseline value from the perspective of causality, and we propose a method to learn the optimal baseline value. Experimental results have demonstrated the effectiveness of our method.

Dataless Knowledge Fusion by Merging Weights of Language Models

- Xisen Jin, Pengxiang Cheng, Daniel Preotiuc-Pietro, Xiang Ren
- abstract@open-review(Poster): 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.

Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval

- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, Ge Yu
- abstract@open-review(Poster): 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 between 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
- abstract@open-review(Poster): Comprehensive studies of synthetic optical flow datasets have attempted to reveal what properties lead to accuracy improvement. However, manually identifying and verifying all such necessary properties are intractable mainly due to the requirement of large-scale trial-and-error experiments with iteratively generating whole synthetic datasets. To tackle this challenge, we propose a differentiable optical flow data generation pipeline and a loss function to drive the pipeline, called DFlow. These enable automatic and efficient synthesis of a dataset effective to a target domain, given a snippet of target data. This distinctiveness is achieved by proposing an efficient data comparison method, where we approximately encode reference sets of data into neural networks and compare the proxy networks instead of explicitly comparing datasets in a sample-wise way. Our experiments show the competitive performance of our DFlow against the prior arts in pre-training. Moreover, the RAFT model pre-trained with DFlow achieves state-of-the-art performance on the Sintel public benchmark in fine-tuning.

Sparse Random Networks for Communication-Efficient Federated Learning

- Berivan Isik, Francesco Pase, Deniz Gunduz, Tsachy Weissman, Zorzi Michele
- abstract@open-review(Poster): 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.

A General Framework For Proving The Equivariant Strong Lottery Ticket Hypothesis

- Damien Ferbach, Christos Tsirigotis, Gauthier Gidel, Joey Bose
- abstract@open-review(Poster): 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 demonstrate 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Graph Domain Adaptation via Theory-Grounded Spectral Regularization

- Yuning You, Tianlong Chen, Zhangyang Wang, Yang Shen
- abstract@open-review(Poster): 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 performances vary and the theoretical foundation has been limited. This paper targets 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 conditional-shift scenario, 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 the numerical agreement from 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 will be fully released upon acceptance.

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
- abstract@open-review(Poster): 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

- Emilio Luz-Ricca, Clare Elizabeth Heinbaugh, Huajie Shao
- abstract@open-review(Poster): 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 autencoder (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.

GReTo: Remedying dynamic graph topology-task discordance via target homophily

- Zhengyang Zhou, qihe huang, Gengyu Lin, Kuo Yang, LEI BAI, Yang Wang
- abstract@open-review(Poster): 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 abovementioned discordance and promote the aggregation capacity. Therefore, a novel 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.

Pareto-Optimal Diagnostic Policy Learning in Clinical Applications via Semi-Model-Based Deep Reinforcement Learning

- Zheng Yu, Yikuan Li, Joseph Chahn Kim, Kaixuan Huang, Yuan Luo, Mengdi Wang
- abstract@open-review(Poster): 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 F_1 score instead of the error rate. However, the F_1 score cannot be written as a cumulative sum of rewards, which invalidates 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 identifies 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.

POPGym: Benchmarking Partially Observable Reinforcement Learning

- Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, Amanda Prorok
- abstract@open-review(Poster): 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 14 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 many possible POMDPs. 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://anonymous.4open.science/r/popgym-51D8/README.md.

Everybody Needs Good Neighbours: An Unsupervised Locality-based Method for Bias Mitigation

- Xudong Han, Timothy Baldwin, Trevor Cohn
- abstract@open-review(Poster): 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.

Particle-based Variational Inference with Preconditioned Functional Gradient Flow

- Hanze Dong, Xi Wang, LIN Yong, Tong Zhang
- abstract@open-review(Poster): 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 remedies the problem by proposing a general framework to obtain tractable functional gradient flow estimates. The functional gradient flow in our framework can be defined by a general functional regularization term that includes the RKHS norm as a special case. We use our framework to propose a new particle-based VI algorithm: preconditioned functional gradient flow (PFG). Compared with SVGD, the proposed method has several advantages: larger function class; greater scalability in large particle-size scenarios; better adaptation to ill-conditioned distributions; provable continuous-time convergence in KL divergence. Non-linear function classes such as neural networks can be incorporated to estimate the gradient flow. Both theory and experiments have shown the effectiveness of our framework.

Learning Locality and Isotropy in Dialogue Modeling

- Han Wu, Haochen Tan, Mingjie Zhan, Gangming Zhao, Shaoqing Lu, Ding Liang, Linqi Song
- abstract@open-review(Poster): 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 across both automatic and human evaluation metrics. More in-depth analyses further confirm the effectiveness of our proposed approach.

Combating Exacerbated Heterogeneity for Robust Decentralized Models

- Jianing Zhu, Jiangchao Yao, Tongliang Liu, quanming yao, Jianliang Xu, Bo Han
- abstract@open-review(Poster): The emerging privacy and security issues in real-world applications motivate us to pursue the adversarially robust federated models. However, the straightforward combination between adversarial training and federated learning in one framework, usually induces the undesired robustness deterioration. We discover that the attribution behind this phenomenon is 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 termed as 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 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 practice 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.

Bidirectional Propagation for Cross-Modal 3D Object Detection

- Yifan Zhang, Qijian Zhang, Junhui Hou, Yixuan Yuan, Guoliang Xing
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 proposed approach, where improvements range from 2% to even over 100% with very limited data. Code and models would be made public.

TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis

- Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, Mingsheng Long
- abstract@open-review(Poster): 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 from 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.

Learning without Prejudices: Continual Unbiased Learning via Benign and Malignant Forgetting

- Myeongho Jeon, Hyoje Lee, Yedarm Seong, Myungjoo Kang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 physical-street-block-based schemes. Although achieving promising results, the hand-crafted schemes severely limit the capabilities of NeRF in large-scale scene modeling. First, it is extremely challenging to manually design a universal scene decomposition rule for different complex scenes, leading to adaptation issues while applying the model to different scenarios. Second, 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, to better model the scene with different sub-networks. Third, the different sub-networks are typically optimized independently, and thus the inconsistency among them cannot be effectively handled during the optimization. To tackle these issues, in this paper, 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.

NORM: Knowledge Distillation via N-to-One Representation Matching

- Xiaolong Liu, Lujun Li, Chao Li, Anbang Yao
- abstract@open-review(Poster): Existing feature distillation methods commonly adopt the One-to-one Representation Matching between each pre-selected teacher-student layer pair. In this paper, we present N-to-One Representation Matching (NORM), a new two-stage knowledge distillation method, which relies on a linear Feature Transform (FT) module. In view of preserving the intact information learnt by the teacher network, during training, our FT module consisting of two linear layers 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 splitting the expanded student representation into N non-overlapping segments having the same number of feature channels as the teacher's, they can be 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 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 CIFAR100 and ImageNet with various teacher-student network pairs show competitive performance of NORM. Code will be released.

Critic Sequential Monte Carlo

- Vasileios Lioutas, Jonathan Wilder Lavington, Justice Sefas, Matthew Niedoba, Yunpeng Liu, Berend Zwartsenberg, Setareh Dabiri, Frank Wood, Adam Scibior
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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.

Deep Learning meets Nonparametric Regression: Are Weight-Decayed DNNs Locally Adaptive?

- Kaiqi Zhang, Yu-Xiang Wang
- abstract@open-review(Poster): 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 ℓp -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
- abstract@open-review(Poster): 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.

Robust Active Distillation

- Cenk Baykal, Khoa Trinh, Fotis Iliopoulos, Gaurav Menghani, Erik Vee
- abstract@open-review(Poster): 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.

Kernel Neural Optimal Transport

- Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev
- abstract@open-review(Poster): We study the Neural Optimal Transport (NOT) algorithm which uses the general optimal transport formulation and learns stochastic transport plans. We show that NOT the weak quadratic cost is doomed to learn fake plans which are not optimal. To resolve this issue, we introduce kernel weak quadratic costs. We show that they provide improved theoretical guarantees and practical performance. We test NOT with kernel costs on the unpaired image-to-image translation task.

SeaFormer: Squeeze-enhanced Axial Transformer for Mobile Semantic Segmentation

- Qiang Wan, Jiachen Lu, Zilong Huang, Gang YU, Li Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning Sparse and Low-Rank Priors for Image Recovery via Iterative Reweighted Least Squares Minimization

- Stamatios Lefkimmiatis, Iaroslav Sergeevich Koshelev
- abstract@open-review(Poster): 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.

Spherical Sliced-Wasserstein

- Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, Minh Tan Pham
- abstract@open-review(Poster): 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.

InPL: Pseudo-labeling the Inliers First for Imbalanced Semi-supervised Learning

- Zhuoran Yu, Yin Li, Yong Jae Lee
- abstract@open-review(Poster): 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.

Maximizing Communication Efficiency for Large-scale Training via 0/1 Adam

- Yucheng Lu, Conglong Li, Minjia Zhang, Christopher De Sa, Yuxiong He
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Strategic Classification on Graphs

- Itay Eilat, Ben Finkelshtein, Chaim Baskin, Nir Rosenfeld
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Learning Symbolic Models for Graph-structured Physical Mechanism

- Hongzhi Shi, Jingtao Ding, Yufan Cao, quanming yao, Li Liu, Yong Li
- abstract@open-review(Poster): 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 of 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.

Priors, Hierarchy, and Information Asymmetry for Skill Transfer in Reinforcement Learning

- Sasha Salter, Kristian Hartikainen, Walter Goodwin, Ingmar Posner
- abstract@open-review(Poster): 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 asymmetric 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 APES, 'Attentive Priors for Expressive and Transferable Skills', 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
- abstract@open-review(Poster): 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.

Causal Representation Learning for Instantaneous and Temporal Effects

- Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, Efstratios Gavves
- abstract@open-review(Poster): 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 temporal sequences with known intervention targets. 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 video datasets, 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
- abstract@open-review(Poster): 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.

CodeT: Code Generation with Generated Tests

- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, Weizhu Chen
- abstract@open-review(Poster): 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 to Generate Columns with Application to Vertex Coloring

- Yuan Sun, Andreas T Ernst, Xiaodong Li, Jake Weiner
- abstract@open-review(Poster): 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.

Towards Real-Time Neural Image Compression With Mask Decay

- Wang Guo-Hua, Jiahao Li, Bin Li, Yan Lu
- abstract@open-review(Poster): 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 network, 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 will be public.

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

- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 policy 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
- abstract@open-review(Poster): 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 a 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 a 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.

Masked Visual-Textual Prediction for Document Image Representation Pretraining

- Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, Jingdong Wang
- abstract@open-review(Poster): In this paper, we present Masked Visual-Textual Prediction for document image representation pretraining, called MaskDoc. It comprises of two self-supervised pretraining tasks: Masked Image Modeling and Masked Language Modeling, based on text region-level image masking. Our approach randomly masks some words or texts and accordingly the corresponding image regions, and the pretraining task is reconstructing the masked image regions as well as the corresponding words. In comparison to masked image modeling which usually predict the image patches or tokens, the encoder pretrained by our approach captures more textual semantics. Compared to the masked multi-modal modeling methods for document image understanding, e.g., LayoutLM and StrucTexT, that need both the image and text inputs, our approach is able to model image-only input, and potentially can deal with more application scenarios free from OCR pre-processing. We demonstrate the effectiveness of MaskDoc on several document image understanding tasks such as image classification, layout analysis, table structure recognition, document OCR, and end-to-end information extraction. Experimental results show that MaskDoc achieves state-of-the-art performance. Our code and models will be released soon.

Plateau in Monotonic Linear Interpolation --- A "Biased" View of Loss Landscape for Deep Networks

- Xiang Wang, Annie N. Wang, Mo Zhou, Rong Ge
- abstract@open-review(Poster): 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.

The KFIoU Loss for Rotated Object Detection

- Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, XIAOPENG ZHANG, Qi Tian
- abstract@open-review(Poster): 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. In contrast, 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 modeing and Kalman filter, which mainly consists of two items. The first term is a scale-insensitive center point loss, which is used to quickly get the center points between bounding boxes closer to assist the second term. In the distance-independent second term, Kalman filter 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 detection. Extensive results on various public datasets (2-D/3-D, aerial/text/face images) with different base detectors show the effectiveness of our approach. The code will be made publicly available.

BrainBERT: Self-supervised representation learning for Intracranial Electrodes

- Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, Andrei Barbu
- abstract@open-review(Poster): 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 approach 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.

General Neural Gauge Fields

- Fangneng Zhan, Lingjie Liu, Adam Kortylewski, Christian Theobalt
- abstract@open-review(Poster): The recent advance of neural radiance fields has significantly pushed the boundary of scene representation learning. Aiming to boost the computational 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 radiance fields. The conversion of measuring systems can be typically dubbed as \emph{gauge transformation}, which is usually some pre-defined mapping functions, e.g., orthogonal projections and spatial hash functions. This begs a question: can we learn the gauge transformation along with the neural scene representations 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 training framework to jointly optimize the gauge transformation and radiance fields. As observing the gauge transformation learning easily suffers from learning collapse, we derive a general regularization mechanism from the principle of information conversation during the gauge transformation. On the strength of the unified neural gauge fields framework, we naturally launch a new type of gauge transformation which suffices to achieve a tradeoff between learning collapse and computation cost.

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
- abstract@open-review(Poster): 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.

Discovering Informative and Robust Positives for Video Domain Adaptation

- Chang Liu, Kunpeng Li, Michael Stopa, Jun Amano, Yun Fu
- abstract@open-review(Poster): 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. Extensive experiments on several cross-domain action recognition benchmarks demonstrate the superiority of our approach over state-of-the-art methods.

Understanding Why Generalized Reweighting Does Not Improve Over ERM

- Runtian Zhai, Chen Dan, J Zico Kolter, Pradeep Kumar Ravikumar
- abstract@open-review(Poster): 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.

Linear Connectivity Reveals Generalization Strategies

- Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, Naomi Saphra
- abstract@open-review(Poster): 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.

Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models

- Meng Liu, Haoran Liu, Shuiwang Ji
- abstract@open-review(Poster): 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 requirement, thereby resulting in difficulties for learning EBMs on high-dimensional data. Motivated from 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.

Composing Ensembles of Pre-trained Models via Iterative Consensus

- Shuang Li, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Igor Mordatch
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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. Comprehensive experiments show that GraphAug outperforms previous graph augmentation methods on various graph classification tasks.

Riemannian Metric Learning via Optimal Transport

- Christopher Scarvelis, Justin Solomon
- abstract@open-review(Poster): 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.

Reliability of CKA as a Similarity Measure in Deep Learning

- MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, Eugene Belilovsky
- abstract@open-review(Poster): 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 recently 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.

Fair Attribute Completion on Graph with Missing Attributes

- Dongliang Guo, Zhixuan Chu, Sheng Li
- abstract@open-review(Poster): 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 missing 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 be applied to any graph and generate fair embeddings 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.

Deep Ranking Ensembles for Hyperparameter Optimization

- Abdus Salam Khazi, Sebastian Pineda Arango, Josif Grabocka
- abstract@open-review(Poster): 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 ranking 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.

Robustness to corruption in pre-trained Bayesian neural networks

- Xi Wang, Laurence Aitchison
- abstract@open-review(Poster): 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.

Weakly-supervised HOI Detection via Prior-guided Bi-level Representation Learning

- Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, Xuming He
- abstract@open-review(Poster): Human object interaction (HOI) detection plays a crucial role in human-centric scene 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

- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 strong baselines and achieves significant improvement over both its Deep RL and EA components.

Deep Ensembles for Graphs with Higher-order Dependencies

- Steven Krieg, William Burgis, Patrick Soga, Nitesh Chawla
- abstract@open-review(Poster): 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.

Towards Understanding Why Mask Reconstruction Pretraining Helps in Downstream Tasks

- Jiachun Pan, Pan Zhou, Shuicheng YAN
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): Category-level articulated object pose estimation aims to estimate a hierarchy of 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.

Thalamus: a brain-inspired algorithm for biologically-plausible continual learning and disentangled representations

- Ali Hummos
- abstract@open-review(Poster): 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

Deep Variational Implicit Processes

- Luis A. Ortega, Simon Rodriguez Santana, Daniel Hernández-Lobato
- abstract@open-review(Poster): 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 are Certifiable Robust Vision Learners

- QuanLin Wu, Hang Ye, Yuntian Gu, Huishuai Zhang, Liwei Wang, Di He
- abstract@open-review(Poster): 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. All model checkpoints and code will be released soon.

Estimating individual treatment effects under unobserved confounding using binary instruments

- Dennis Frauen, Stefan Feuerriegel
- abstract@open-review(Poster): Estimating individual treatment effects (ITEs) 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 compliance is used as binary IV). In this paper, we propose a novel, multiply robust machine learning framework, called MRIV, for estimating ITEs using binary IVs and thus yield an unbiased ITE estimator. Different from previous work for binary IVs, our framework estimates the ITE directly via a pseudo outcome regression. (1)~We provide a theoretical analysis where we show that our framework yields multiple robust convergence rates: our ITE 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 ITE estimation. (3)~We build upon our theoretical results and propose a tailored deep neural network architecture called MRIV-Net for ITE 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 ITEs in the binary IV setting.

Approximate Bayesian Inference with Stein Functional Variational Gradient Descent

- Tobias Pielok, Bernd Bischl, David Rügamer
- abstract@open-review(Poster): 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.

SCoMoE: Efficient Mixtures of Experts with Structured Communication

- zhiyuan zeng, Deyi Xiong
- abstract@open-review(Poster): Mixture-of-Experts (MoE) models are promising architectures for massive 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 massive bilingual and 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.

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
- abstract@open-review(Poster): 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 selecting explanatory input features, which follow either additive or instance-wise directions. Additive methods exploit local neighbourhoods to learn instance-specific explainers sequentially. The process is thus inefficient and susceptible to poorly-conditioned samples. Meanwhile, instance-wise methods directly optimize local feature distributions in a global training framework, thereby being capable of leveraging global information from other inputs. However, they can only interpret single-class predictions and 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.

LDMIC: Learning-based Distributed Multi-view Image Coding

- Xinjie Zhang, Jiawei Shao, Jun Zhang
- abstract@open-review(Poster): 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 constraint 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 epipolar geometry relations between images. Experimental results show that LDMIC significantly outperforms both traditional and learning-based MIC methods while enjoying fast encoding speed.

Sound Randomized Smoothing in Floating-Point Arithmetic

- Vaclav Voracek, Matthias Hein
- abstract@open-review(Poster): 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.

Collaborative Pure Exploration in Kernel Bandit

- Yihan Du, Wei Chen, Yuko Kuroki, Longbo Huang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Test-Time Robust Personalization for Federated Learning

- Liangze Jiang, Tao Lin
- abstract@open-review(Poster): 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 step further: robustifying FL models under 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 performance and robustness of personalized FL methods during deployment.

Learning to Linearize Deep Neural Networks for Secure and Efficient Private Inference

- Souvik Kundu, Shunlin Lu, Yuke Zhang, Jacqueline Tiffany Liu, Peter Anthony Beerel
- abstract@open-review(Poster): 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.

Meta Knowledge Condensation for Federated Learning

- Ping Liu, Xin Yu, Joey Tianyi Zhou
- abstract@open-review(Poster): 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).

Masked Frequency Modeling for Self-Supervised Visual Pre-Training

- Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, Chen Change Loy
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 relies 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 the selection of in-context examples.

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
- abstract@open-review(Poster): 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 retrained 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 implicitly. 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.

Bort: Towards Explainable Neural Networks with Bounded Orthogonal Constraint

- Borui Zhang, Wenzhao Zheng, Jie Zhou, Jiwen Lu
- abstract@open-review(Poster): 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 model's comprehensibility and transparency. 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 transparency. We perform reconstruction and backtracking on the model representations optimized by Bort and observe an evident 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.

The Power of Regularization in Solving Extensive-Form Games

- Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, Kaiqing Zhang
- abstract@open-review(Poster): 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 without the uniqueness assumption of the Nash equilibrium (NE). Moreover, regularized dilated optimistic multiplicative weights update (\texttt{Reg-DOMWU}), an instance of \texttt{Reg-DOMD}, further enjoys the $\tilde O(1/T)$ last-iterate convergence rate of the distance to the set of NE. This addresses an open question of whether iterate convergence could be obtained for OMWU-type algorithms with constant stepsizes, without the unique NE assumption in both the EFG and normal-form game literature. 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.

MLPInit: Embarrassingly Simple GNN Training Acceleration with MLP Initialization

- Xiaotian Han, Tong Zhao, Yozen Liu, Xia Hu, Neil Shah
- abstract@open-review(Poster): 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) whose weights can be made identical, 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 for graph data, 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). Most importantly, MLPInit is extremely simple to implement and can be flexibly used as a plug-and-play initialization method for message passing-based GNNs

Progressive 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
- abstract@open-review(Poster): Masked Image Modeling (MIM) methods are 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 in MIM. This redundancy is neglected by existing methods and causes non-negligible overheads in computation and storage that do not necessarily benefit self-supervised learning. In this paper, we present a novel approach named Progressive Compressed AutoEncoder (PCAE) to address this problem by progressively compacting tokens and retaining the least necessary information for representation. In particular, we propose to mitigate the performance degradation caused by token reduction through exploiting the vision transformer to leak information from discarded tokens to the retained ones. Besides, we also propose the progressive discarding strategy to achieve a better trade-off between performance and efficiency. Identifying redundant tokens plays a key role in redundancy reduction. We resolve this issue using a simple yet effective criterion, i.e., we identify redundant tokens according to their similarity to the mean of token sequence. Thanks to the flexible strategy, PCAE can be employed for both pre-training and downstream fine-tuning and, consequently, reduces the computing overhead non-trivially throughout the training pipeline. Experiments show that PCAE achieves comparable performance while at most accelerates 1.9 times throughput compared with MAE for self-supervised learning, and accelerates 15\%-57\% throughput while the performance drop is within 0.6\% for downstream classification.

S-NeRF: Neural Radiance Fields for Street Views

- Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, Li Zhang
- abstract@open-review(Poster): 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 collected 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.

Cycle-consistent Masked AutoEncoder for Unsupervised Domain Generalization

- Haiyang Yang, SHIXIANG TANG, Xiaotong Li, Feng Zhu, Yizhou Wang, Meilin Chen, LEI BAI, Rui Zhao, Wanli Ouyang
- abstract@open-review(Poster): 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. Codes shall be released upon acceptance.

CFlowNets: Continuous control with Generative Flow Networks

- Yinchuan Li, Shuang Luo, Haozhi Wang, Jianye HAO
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Deep generative models such as GANs and normalizing flows 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 an inversion process, particularly in the presence of data noise and inaccurate forward models. In such cases, deep generative models are ineffective in attaining high-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 tested and validated 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, and achieved 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

- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 feature 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 modelling 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. Code will be available.

Causal Balancing for Domain Generalization

- Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, William Yang Wang
- abstract@open-review(Poster): 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.

Towards Addressing Label Skews in One-Shot Federated Learning

- Yiqun Diao, Qinbin Li, Bingsheng He
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Recently, generalization on out-of-distribution (OOD) data with correlation shift has attracted great attention. 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 conditionally independent models of spurious attributes are OOD generalizable. Based on this, a metric Conditional Spurious Variation (CSV) which controls 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 objective 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.

Relaxed Combinatorial Optimization Networks with Self-Supervision: Theoretical and Empirical Notes on the Cardinality-Constrained Case

- Runzhong Wang, Li Shen, Yiting Chen, Xiaokang Yang, Dacheng Tao, Junchi Yan
- abstract@open-review(Poster): Self-supervised neural networks for combinatorial optimization (CO) handle non-differentiable constraints via relaxation. Despite their superiority in efficiency, one possible limitation is that these methods often put the constraints as soft penalty terms in the learning objective, and the degree of constraint-violation usually cannot be accurately or directly modulated. In this paper, we aim to develop a new paradigm to solve the CO problem by incorporating the constraints into the network architecture and computational operators, which is a more natural learning pipeline and decouples the constraint violation penalty from the raw objective optimization. Seeing such a paradigm may be rather general such that there only exist perturbation-based blackbox differentiable learning methods as generic solvers in literature, here we consider the commonly used cardinality constraints which in fact can incorporate many existing CO problem instances as its special cases. Specifically, the cardinality constraints are encoded by a differentiable optimal transport layer. We theoretically characterize the constraint-violations of two variants of our architecture (w.r.t. existing CO network whose constraint-violation is non-controlled), and we further show that their empirical performances are in line with our theoretical results. On self-supervised learning of pure CO problems on synthetic and real-world data, our networks surpass the state-of-the-art CO network, and are comparable to Gurobi and can sometimes even surpass. Our general paradigm also enables the application of end-to-end predictive portfolio optimization on real-world asset price data, improving the Sharpe ratio from 1.1 to 2.1 with a predict-then-optimize paradigm with LSTM+Gurobi.

Block and Subword-Scaling Floating-Point (BSFP) : An Efficient Non-Uniform Quantization For Low Precision Inference

- Yun-Chen Lo, Tse-Kuang Lee, Ren-Shuo Liu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 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 the 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. We demonstrate that self-supervised AT can attain even better robustness than supervised AT for the first time.

Semi-supervised Community Detection via Structural Similarity Metrics

- Yicong Jiang, Tracy Ke
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 the 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
- abstract@open-review(Poster): 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.

Automating Nearest Neighbor Search Configuration with Constrained Optimization

- Philip Sun, Ruiqi Guo, Sanjiv Kumar
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Effective Self-supervised Pre-training on Low-compute networks without Distillation

- Fuwen Tan, Fatemeh Sadat Saleh, Brais Martinez
- abstract@open-review(Poster): 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, and we propose an effective training strategy that achieves the new state-of-the-art for SSL on low-compute networks despite not using KD at all. In particular, we start by empirically observing that the use of local views can have a dramatic impact on the effectiveness of the SSL method. This hints at view sampling being the performance bottleneck 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 performance by a wide margin across SSL methods (e.g. MoCo-v2, SwAV or DINO), across low-size networks (convolution-based networks, e.g. MobileNetV2, ResNet18, ResNet34 and vision transformer, e.g. ViT-Ti), and across tasks (linear probe, object detection, instance segmentation, and semi-supervised learning). Our best models establish a new state-of-the-art for SSL methods on low-compute networks across all standard benchmarks despite not using a KD loss term.

CoRTX: Contrastive Framework for Real-time Explanation

- Yu-Neng Chuang, Guanchu Wang, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanting Cai, Xia Hu
- abstract@open-review(Poster): 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.

OTOv2: Automatic, Generic, User-Friendly

- Tianyi Chen, Luming Liang, Tianyu DING, Zhihui Zhu, Ilya Zharkov
- abstract@open-review(Poster): Only-Train-Once (OTOv1) is recently proposed to drastically simplify and automate the complicated multi-stage procedure for model compression via structured pruning. However, its automation relies on manually conducting zero-invariant groups (ZIGs) partition and slimmer model construction for specific DNNs beforehand, which necessitates numerous engineering efforts and domain-knowledge and prevents its wider applications onto general scenarios. We propose the second generation of Only-Train-Once (OTOv2), which trains and compresses an arbitrary DNN only once from scratch to produce a more compact model with competitive performance without fine-tuning. OTOv2 is automated and pluggable into various deep learning applications, and requires almost minimal engineering efforts from the users. Methodologically, OTOv2 proposes two revolutionary improvements: (i) Autonomy: automatically partitions ZIGs and constructs compressed model for arbitrary DNNs; 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, DenseNet and StackedUnets, the majority of which cannot be handled by OTOv1 without extensive handcrafting. Together with benchmark datasets including CIFAR10/100, Fashion-MNIST, SVNH and ImageNet, the effectiveness of OTOv2 is validated by achieving competitive or even better results than the state-of-the-arts.

Filter-Recovery Network for Multi-Speaker Audio-Visual Speech Separation

- Haoyue Cheng, Zhaoyang Liu, Wayne Wu, Limin Wang
- abstract@open-review(Poster): We aim at audio-visual speech separation task. Given the face information for each speaker, the goal is to separate the corresponding speech in the speech mixture. Existing works are designed for a controlled setting with a fixed number of speakers, mostly 2 or 3 speakers, which is not easily scalable in practical application. To deal with this, we focus on separating voices for variable number of speakers with a single model, and build concrete mixture test sets for a fair comparison. There are two prominent issues in complex multi-speaker separation results: 1) There exists some noisy voice pieces belong to other speakers; 2) Part of the target speech is missing. To deal with these, we propose a valid method BFRNet, including a basic audio-visual speech separator and a Filter-Recovery Network (FRNet). The FRNet filters the noisy speech and recovery the missing parts for the output of the basic separator. Our method achieves the state-of-the-art results on audio-visual speech separation datasets. Besides, we apply the FRNet to other methods and achieve general performance improvements, which proves the effectiveness of the proposed FRNet.

Can discrete information extraction prompts generalize across language models?

- Nathanaël Carraz Rakotonirina, Roberto Dessi, Fabio Petroni, Sebastian Riedel, Marco Baroni
- abstract@open-review(Poster): 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.

A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta.

- Maksim Velikanov, Denis Kuznedelev, Dmitry Yarotsky
- abstract@open-review(Poster): 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.

Spike Calibration: Bridging the Gap between ANNs and SNNs in ANN-SNN Conversion

- Zecheng Hao, Jianhao Ding, Tong Bu, Tiejun Huang, Zhaofei Yu
- abstract@open-review(Poster): Spiking Neural Networks (SNNs) have attracted great attention due to the distinctive characteristics of low power consumption and temporal information processing. ANN-SNN conversion, as the most commonly used method, can make converted SNNs achieve comparable performance as ANNs on large-scale datasets. However, the performance degrades severely under low time-steps, which hampers the practical applications of SNNs on neuromorphic chips. In this paper, instead of evaluating different conversion errors and then eliminating these errors, we define offset spike to measure the deviation degree of actual and

desired firing rates of SNNs. We make a detailed analysis of offset spike and point out that the case of firing one more (or less) spike is the main reason for conversion error. Based on this, we propose an optimization strategy based on shifting initial membrane potential and theoretically prove the corresponding optimal shifting distance to calibrate the spike. In addition, we also note that our method has a unique iterative property to further reduce conversion error. 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 top-1 accuracy of 67.12% on ImageNet with 6 time-steps. To the best of our knowledge, this is the first time ANN-SNN conversion can simultaneously achieve high accuracy and ultra-low latency on the complex dataset.

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
- abstract@open-review(Poster): Recent 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. On the other hand, various trainable calibration measures have been proposed recently to incorporate the calibration objective loss directly into the training process. However, these methods all incorporate internal hyperparameters introduced in the process of obtaining a differential calibration measure. Consequently, the performance of these calibration objectives relies on tuning these hyperparameters, incurring more computational cost 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 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 hyperparmeter tuning, (2) ESD yields the best calibrated results compared with previous approaches, (3) show that ESD drastically improve the computational cost required for calibration during training due to the absence of internal hyperparamater. Code will be publicly available.

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
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Contextual Convolutional Networks

- Shuxian Liang, Xu Shen, Tongliang Liu, Xian-Sheng Hua
- abstract@open-review(Poster): 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 will be released in the final version.

Statistical Inference for Fisher Market Equilibrium

- Luofeng Liao, Yuan Gao, Christian Kroer
- abstract@open-review(Poster): 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.

Scenario-based Question Answering with Interacting Contextual Properties

- Haitian Sun, William W. Cohen, Ruslan Salakhutdinov
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

LPT: Long-tailed Prompt Tuning for Image Classification

- Bowen Dong, Pan Zhou, Shuicheng Yan, Wangmeng Zuo
- abstract@open-review(Poster): 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 pretrained 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.

Digging into Backbone Design on Face Detection

- Yang Liu, Fei Wang, Lei Shang, Jiankang Deng, Baigui Sun, Xuansong Xie
- abstract@open-review(Poster): 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. However, 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 costs in detection frameworks. Considering the intrinsic design property 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.

Towards Smooth Video Composition

- Qihang Zhang, Ceyuan Yang, Yujun Shen, Yinghao Xu, Bolei Zhou
- abstract@open-review(Poster): 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 will be made publicly available.

DiffMimic: Efficient Motion Mimicking with Differentiable Physics

- Jiawei Ren, Cunjun Yu, Siwei Chen, Xiao Ma, Liang Pan, Ziwei Liu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Non-determinism in deep learning and the consequential randomness and variability in performance evaluation has spawned attempts to foster reproducibility of SOTA benchmark results by sharing data, code, and meta-parameter settings. In this paper, we propose to shift from the goal of duplicating a SOTA training result without any changes to a new type of reproducibility called inferential reproducibility that treats performance variation depending on data characteristics, meta-parameter settings, and their interactions as an inherent and interesting feature of non-deterministic deep learning, not as a bug that needs to be resolved. We propose to answer questions of inferential reproducibility by classical statistical methods: We show how to design a linear mixed effects model (LMEM) to analyze performance evaluation scores of machine learning algorithms, and to conduct statistical inference on the interpretable parameters of this model with a generalized likelihood ratio test (GLRT). This approach allows us to efficiently assess statistical significance of performance differences between models by simultaneously acknowledging for variability in meta-parameters and data. Furthermore, performance differences conditional on data properties can be assessed, and

a variance component analysis (VCA) can be performed to reveal the contribution of meta-parameters to overall variance. Lastly, a reliability coefficient can be computed to assess the general robustness of the model. Code (R and Python) and sample applications of our tools are publicly available.

Knowledge Distillation based Degradation Estimation for Blind Super-Resolution

- Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, Luc Van Gool
- abstract@open-review(Poster): 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 settings. 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.

Graph Contrastive Learning for Skeleton-based Action Recognition

- Xiaohu Huang, Hao Zhou, Bin Feng, Xinggang Wang, Wenyu Liu, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Spikformer: When Spiking Neural Network Meets Transformer

- Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, Li Yuan
- abstract@open-review(Poster): 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.

Multimodal Analogical Reasoning over Knowledge Graphs

- Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, Huajun Chen
- abstract@open-review(Poster): 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 a knowledge graph, 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.

MECTA: Memory-Economic Continual Test-Time Model Adaptation

- Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, Michael Spranger
- abstract@open-review(Poster): 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 8.5% with constrained memory and significantly reduces the memory cots of ResNet50 on ImageNet by at least 70% without sacrificing accuracy. Our code will be published upon acceptance.

Interpretability with full complexity by constraining feature information

- Kieran A Murphy, Danielle Bassett
- abstract@open-review(Poster): Interpretability is a pressing issue for machine learning. Common approaches to interpretable machine learning constrain interactions between features of the input, rendering comprehensible the effects of features on a model's output at the expense of model complexity. We approach interpretability from a different angle: constrain the information about the features utilized by the model, without any restrictions on the complexity of the model. Borrowing from information theory, we use the Distributed Information Bottleneck to find optimal compressions of each feature that maximally preserve information about the output. The learned information allocation, by feature and by feature value, is a rich source of interpretability well-suited to 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 and range from the uninformative to the most performant model obtainable. Information is allocated to features by their relevance to the output, tackling the problem of feature selection with inclusion/exclusion existing on a learned continuum. The optimal compression of each feature---at every stage of approximation---allows fine-grained inspection of how feature values are similar or distinct with regards to the prediction. We develop a framework for extracting insight from the spectrum of approximate models and demonstrate its utility on a range of tabular datasets.

What shapes the loss landscape of self supervised learning?

- Liu Ziyin, Ekdeep Singh Lubana, Masahito Ueda, Hidenori Tanaka
- abstract@open-review(Poster): 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 provide answers to these questions by thoroughly analyzing SSL loss landscapes for a linear model. We derive an analytically tractable theory of SSL landscape and show that it accurately captures an array of collapse phenomena and identifies their causes. 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.

Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies

- Rui Yuan, Simon Shaolei Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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 two new algorithms, LinPEVI-ADV+ and LinPMVI-ADV+, for single-agent MDPs and two-player zero-sum Markov games (MGs), respectively. The proposed algorithms establish pessimism in a variance-reduction manner via reference-advantage decomposition and variance-reweighted ridge regression. Theoretical analysis demonstrates that they can match the performance lower bounds up to logarithmic factors. We also establish new performance lower bounds for MDPs and MGs, which tighten the existing results, to demonstrate the nearly minimax optimality of the proposed algorithms. As a byproduct, equipped with the techniques developed in this paper, we can further improve the suboptimality bound when the feature vector set is finite. 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.

Conditional Positional Encodings for Vision Transformers

- Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Chunhua Shen
- abstract@open-review(Poster): 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.

ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills

- Jiayuan Gu, Fanbo Xiang, Zhan Ling, Xinyue Wei, Xiqiang Liu, Xuanlin Li, Rui Chen, Stone Tao, Tongzhou Mu, Pengwei Xie, Yunchao Yao, Yihe Tang, Xiaodi Yuan, Zhiao Huang, Hao Su
- abstract@open-review(Poster): 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 (e.g., stationary/mobile-base, single/dual-arm, rigid/soft-body). 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, which 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 will open-source all codes of our benchmark (simulator, environments, and baselines) and host an online challenge open to interdisciplinary researchers.

Twofer: Tackling Continual Domain Shift with Simultaneous Domain Generalization and Adaptation

- Chenxi Liu, Lixu Wang, Lingjuan Lyu, Chen Sun, Xiao Wang, Qi Zhu
- abstract@open-review(Poster): 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, especially when domain shifts happen suddenly and momentarily. 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 or DG in tackling continual domain shifts, we propose Twofer, a framework that simultaneously achieves target domain generalization (TDG), target domain adaptation (TDA), and forgetting alleviation (FA). Twofer 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 Domain Net datasets demonstrate that Twofer substantially outperforms state-of-the-art works in Continual DA, Source-Free DA, Test-Time/Online DA, Single DG, Multiple DG, and Unified DA&DG. We envision this work as a significant milestone in tackling continual data domain shifts, with improved performance across target domain generalization, adaptation, and forgetting alleviation abilities.

ModelAngelo: Automated Model Building for Cryo-EM Maps

- Kiarash Jamali, Dari Kimanius, Sjors HW Scheres
- abstract@open-review(Poster): 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.

Distilling Cognitive Backdoor Patterns within an Image

- Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey
- abstract@open-review(Poster): 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.

Revocable Deep Reinforcement Learning with Affinity Regularization for Outlier-Robust Graph Matching

- Chang Liu, Zetian Jiang, Runzhong Wang, Lingxiao Huang, Pinyan Lu, Junchi Yan
- abstract@open-review(Poster): Graph matching (GM) has been a building block in various areas including computer vision and pattern recognition. Despite the recent impressive progress, existing deep GM methods often have 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 task. 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: 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.

One Transformer Can Understand Both 2D & 3D Molecular Data

- Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, Di He
- abstract@open-review(Poster): 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 \url{https://anonymous}.

Mind the Gap: Offline Policy Optimizaiton for Imperfect Rewards

- Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, Qing-Shan Jia, Ya-Qin Zhang
- abstract@open-review(Poster): Reward function is essential in reinforcement learning (RL), serving as the guiding signal to incentivize an agent to solve a given task. However, reward function is notoriously difficult to design or even approximate. In many cases, only a sub-par reward function can be obtained, sometimes even with zero reward signal, which often inflicts substantial performance loss or stringent restrictive requirements on expert demonstrations. 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 state-action visitation distribution matching w.r.t. a small set of expert data; and the lower layer solves a pessimistic RL problem with the corrected rewards. By exploiting the duality of the lower level problem, 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, as well as retrieving useful information from biased rewards.

Learning to Compose Soft Prompts for Compositional Zero-Shot Learning

- Nihal V. Nayak, Peilin Yu, Stephen Bach
- abstract@open-review(Poster): 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.

SQA3D: Situated Question Answering in 3D Scenes

- Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, Siyuan Huang
- abstract@open-review(Poster): 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.

Empowering Networks With Scale and Rotation Equivariance Using A Similarity Convolution

- Zikai Sun, Thierry Blu
- abstract@open-review(Poster): The translational equivariant nature of CNN is a reason for its great success in the field of computer vision. However, networks do not enjoy more general equivariance properties such as rotation or scaling. This limits the generalization performance of the network. In this paper, we devise a method that provides networks with equivariance with respect to translation, rotation, and scaling simultaneously. We define a convolution-like operation and ensure equivariance based on our proposed scalable Fourier-Argand representation. The method has similar efficiency as a traditional network, since it hardly introduces any additional learnable parameters and does not rely on group theory. We verified the quality of our approach in the image classification task, demonstrating the robustness and the generalization ability to both scaled and rotated inputs.

Robust and Controllable Object-Centric Learning through Energy-based Models

- Ruixiang ZHANG, Tong Che, Boris Ivanovic, Renhao Wang, Marco Pavone, Yoshua Bengio, Liam Paull
- abstract@open-review(Poster): 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.

Topology-aware robust optimization

- Fengchun Qiao, Xi Peng
- abstract@open-review(Poster): 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 that the learned topology is highly explainable and consistent with human knowledge and scientific plausibility.

EAGLE: Large-scale Learning of Turbulent Fluid Dynamics with Mesh Transformers

- Steeven JANNY, Aurélien Bénéteau, Madiha Nadri, Julie Digne, Nicolas THOME, Christian Wolf
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

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
- abstract@open-review(Poster): 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. Specifically, 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.

Revisiting the Entropy Semiring for Neural Speech Recognition

- Oscar Chang, Dongseong Hwang, Olivier Siohan
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): Over 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 which are prone to get trapped in local optimal points. In order to 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.

Measuring axiomatic identifiability of counterfactual image models

- Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, Ben Glocker
- abstract@open-review(Poster): 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 identifiability 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 models and to provide insight into a model's identifiability shortcomings and trade-offs.

Alternating Differentiation for Optimization Layers

- Haixiang Sun, Ye Shi, Jingya Wang, Hoang Duong Tuan, H. Vincent Poor, Dacheng Tao
- abstract@open-review(Poster): 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 and thus significantly increases the computational speed of implicit differentiation. Further, we present the computational complexity of the forward and backward pass of Alt-Diff and show that Alt-Diff enjoys quadratic computational complexity in the backward pass. Another notable difference between Alt-Diff and state-of-the-arts is that Alt-Diff can be truncated for the optimization layer. We theoretically show that: 1) Alt-Diff can converge to consistent gradients obtained by differentiating KKT conditions; 2) the error between the gradient obtained by the truncated Alt-Diff and by differentiating KKT conditions is upper bounded by the same order of variables' truncation error. Therefore, Alt-Diff can be truncated to further increases computational speed without sacrificing much accuracy. A series of comprehensive experiments demonstrate that Alt-Diff yields results comparable to the state-of-the-arts in far less time.

Out-of-distribution Detection with Implicit Outlier Transformation

- Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye HAO, Bo Han
- abstract@open-review(Poster): 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---searching 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.

Extracting Robust Models with Uncertain Examples

- Guanlin Li, Guowen Xu, Shangwei Guo, Han Qiu, Jiwei Li, Tianwei Zhang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

E-CRF: Embedded Conditional Random Field for Boundary-caused Class Weights Confusion in Semantic Segmentation

- Jie Zhu, Huabin Huang, Banghuai Li, Leye Wang
- abstract@open-review(Poster): 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. Code will be available.

Sample Complexity of Nonparametric Off-Policy Evaluation on Low-Dimensional Manifolds using Deep Networks

- Xiang Ji, Minshuo Chen, Mengdi Wang, Tuo Zhao
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

On The Inadequacy of Optimizing Alignment and Uniformity in Contrastive Learning of Sentence Representations

- Zhijie Nie, Richong Zhang, Yongyi Mao
- abstract@open-review(Poster): 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.

Volumetric Optimal Transportation by Fast Fourier Transform

- Na Lei, DONGSHENG An, Min Zhang, Xiaoyin Xu, David Gu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Hierarchical Relational Learning for Few-Shot Knowledge Graph Completion

- Han Wu, Jie Yin, Bala Rajaratnam, Jianyuan Guo
- abstract@open-review(Poster): 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.

Function-Consistent Feature Distillation

- Dongyang Liu, Meina Kan, Shiguang Shan, Xilin CHEN
- abstract@open-review(Poster): As a commonly used technique in model compression of deep neural networks, feature distillation makes the student model mimic the intermediate features of the teacher model, in hopes that the underlying knowledge in the features could provide extra guidance to the student. 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 the lateral parts of the network will read, decode, and process them. Therefore, we propose 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 lateral 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. Codes will be publicly released soon.

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
- abstract@open-review(Poster): 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.

MCAL: Minimum Cost Human-Machine Active Labeling

- Hang Qiu, Krishna Chintalapudi, Ramesh Govindan
- abstract@open-review(Poster): Today, groundtruth generation relies on datasets annotated by cloud-based annotation services. These 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 dataset, and is always cheaper than the cheapest competing strategy.

Learnable Topological Features For Phylogenetic Inference via Graph Neural Networks

- Cheng Zhang
- abstract@open-review(Poster): 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.

Fairness-aware Contrastive Learning with Partially Annotated Sensitive Attributes

- Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, Jun Xiao
- abstract@open-review(Poster): 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.

Rotamer Density Estimators are Unsupervised Learners of the Effect of Mutations on Protein-Protein Interaction

- Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, Jianzhu Ma
- abstract@open-review(Poster): Protein-protein interactions play a fundamental role in a broad range of biological processes. Predicting the effect of amino acid mutations on binding is crucial to protein engineering. Traditional biophysical and statistical methods have dominated the area for years, but they depend heavily on expert prior and face the trade-off between efficiency and accuracy. Recent success in deep learning for proteins has made data-driven approaches more appealing than ever. Nevertheless, the major challenge is the scarcity of experimental mutational data annotated with the change in binding affinity. In this work, we demonstrate that mutational effects on binding can be predicted by the change in conformational flexibility of the protein-protein interface. We propose a flow-based generative model to estimate the probability distribution of conformation (named Rotamer Density Estimator, RDE) and use entropy as the measure of flexibility. The model is trained solely with protein structures and does not require the supervision of the experimental values of changes in binding affinities. Further, the unsupervised representations extracted by the model can be used for prediction even more accurately using simple downstream neural networks. The proposed method outperforms empirical energy functions and other machine learning-based approaches.

Dilated convolution with learnable spacings

- Ismail Khalfaoui Hassani, Thomas Pellegrini, Timothée Masquelier

- abstract@open-review(Poster): 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 with both ResNet50 and ConvNeXt.

PatchDCT: Patch Refinement for High Quality Instance Segmentation

- Qinrou Wen, Jirui Yang, Xue Yang, Kewei Liang
- abstract@open-review(Poster): 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, and the code will be made publicly available.

ChiroDiff: Modelling chirographic data with Diffusion Models

- Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, Yi-Zhe Song
- abstract@open-review(Poster): 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, controlled abstraction are also possible with this model-class. We perform quantitative and qualitative evaluation of our framework on relevant datasets (VectorMNIST, KanjiVG, Quick, Draw! etc) and found it to be better or on par with competing approaches.

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
- abstract@open-review(Poster): 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.

Cross-Level Distillation and Feature Denoising for Cross-Domain Few-Shot Classification

- Hao ZHENG, Runqi Wang, Jianzhuang Liu, Asako Kanezaki
- abstract@open-review(Poster): 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 soon.

DELTA: DEGRADATION-FREE FULLY TEST-TIME ADAPTATION

- Bowen Zhao, Chen Chen, Shu-Tao Xia
- abstract@open-review(Poster): 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

- abstract@open-review(Poster): 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.

KNN-Diffusion: Image Generation via Large-Scale Retrieval

- Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, Yaniv Taigman
- abstract@open-review(Poster): 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 without any text, (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 (without paired text data).

Decompose to Generalize: Species-Generalized Animal Pose Estimation

- Guangrui Li, Yifan Sun, Zongxin Yang, Yi Yang
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 blAck-box modeLs} (IDEAL), which aims to query-efficiently learn from black-box model APIs to train a good student without any real data. % a small number of queries. 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. Our code will be published upon acceptance.

Trainability Preserving Neural Pruning

- Huan Wang, Yun Fu
- abstract@open-review(Poster): Many recent pruning works show trainability plays a critical role in network structured pruning -- unattended broken trainability can lead to severe under-performance and unintentionally amplify the effect of finetuning learning rate, resulting in biased (or even misinterpreted) benchmark results. In this paper, we present trainability preserving pruning (TPP), a scalable method to preserve network trainability against pruning, aiming for improved pruning performance. TPP regularizes 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. Empirically, TPP performs on par with the ground-truth trainability recovery method on linear MLP networks. On non-linear networks (ResNet56/VGG19 on CIFAR10/100), our TPP outperforms the other counterpart schemes by an obvious margin. Moreover, extensive results on ImageNet with ResNets show TPP consistently performs more favorably against other top-performing structured pruning approaches.

DrML: Diagnosing and Rectifying Vision Models using Language

- Yuhui Zhang, Jeff Z. HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, Serena Yeung
- abstract@open-review(Poster): 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, DrML, 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 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
- abstract@open-review(Poster): 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.

DropIT: Dropping Intermediate Tensors for Memory-Efficient DNN Training

- Joya Chen, Kai Xu, Yuhui Wang, Yifei Cheng, Angela Yao
- abstract@open-review(Poster): 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). Our code and models are available at https://anonymous.4open.science/r/dropit-iclr177submission.

A Unified Framework of Soft Threshold Pruning

- Yanqi Chen, Zhaofei Yu, Wei Fang, Zhengyu Ma, Xiawu Zheng, Yonghong Tian
- abstract@open-review(Poster): Soft threshold pruning is among the cutting-edge pruning methods with state-of-the-art performance. However, previous methods either aimlessly perform searching on the threshold scheduler or simply train the threshold, 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 explained as a specific arrangement style of 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 in the equivalent optimization problem stable and corresponds to a consistent objective function and can be, in principle, applied to sparsify any mathematical model that includes parameters 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. It further evolves into a family of novel pruning methods, including sparsify-during-training, early pruning, and pruning at initialization via analysis based on our framework.

TaskPrompter: Spatial-Channel Multi-Task Prompting for Dense Scene Understanding

- Hanrong Ye, Dan Xu
- abstract@open-review(Poster): 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 jointly 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 establish 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 superiority of the proposed framework and TaskPrompter establishes significant state-of-the-art performances on multi-task dense predictions. Code and models will be made publicly available.

Learning Domain-Agnostic Representation for Disease Diagnosis

- Churan Wang, Jing Li, Xinwei Sun, Fandong Zhang, Yizhou Yu, Yizhou Wang
- abstract@open-review(Poster): 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.

Logical Entity Representation in Knowledge-Graphs for Differentiable Rule Learning

- Chi Han, Qizheng He, Charles Yu, Xinya Du, Hanghang Tong, Heng Ji
- abstract@open-review(Poster): 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.

BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection

- Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, Feng Zhao
- abstract@open-review(Poster): 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.

Interpretable Single/Multi-label Text Classification with Unsupervised Constituent-label alignments

- Xiang Hu, XinYu KONG, Kewei Tu
- abstract@open-review(Poster): 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 structured language models. Specifically, we propose a novel label extraction framework based on binary syntax trees. Both the structures and intermediate representations of the trees are learned using a pretrained neural network in an unsupervised manner. Inference and learning is made efficient using dynamic programming over tree structures. Our experiments demonstrate that our approach could achieve good prediction results in single/multi-label text classification and have explicit and inherent constituent-level interpretability.

Suppressing the Heterogeneity: A Strong Feature Extractor for Few-shot Segmentation

- Zhengdong Hu, Yifan Sun, Yi Yang
- abstract@open-review(Poster): 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.

Achieve the Minimum Width of Neural Networks for Universal Approximation

- Yongqiang Cai
- abstract@open-review(Poster): 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: Horizonal Box Annotation is All You Need for Oriented Object Detection

- Xue Yang, Gefan Zhang, Wentong Li, Yue Zhou, Xuehui Wang, Junchi Yan
- abstract@open-review(Poster): 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, and even surpasses them in some cases. Source code will be made publicly available.

Pushing the Limits of Fewshot Anomaly Detection in Industry Vision: Graphcore

- Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, Feng Zheng
- abstract@open-review(Poster): In the area of fewshot anomaly detection (FSAD), efficient visual feature plays an essential role in memory bank 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 in industrial-based FSAD. Specifically, we utilize of graph representation in FSAD and provide a novel visual isometric invariant feature (VIIF) as 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 M by a large amount. Besides, we provide a novel model GraphCore via VIIFs that can fast implement unsupervised FSAD training and can improve the performance of anomaly detection. A comprehensive evaluation is provided for comparing GraphCore and other SOTA anomaly detection models under our proposed fewshot 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.

Representation Learning for Low-rank General-sum Markov Games

- Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Zihan Ding, Chi Jin, Mengdi Wang
- abstract@open-review(Poster): 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 widely used deep RL baseline, DQN with fictitious play.

Surgical Fine-Tuning Improves Adaptation to Distribution Shifts

- Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, Chelsea Finn
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

On amortizing convex conjugates for optimal transport

- Brandon Amos
- abstract@open-review(Poster): 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. I show that combining an amortized approximations to the conjugate with an exact solver is computationally easy. This combination 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. To attain these results, I have also implemented a new parallel Armijo line search for L-BFGS that runs in ~3% of the time as Jax's default sequential Wolfe line search. All of the baselines, methods, and solvers considered in this paper are also available as part of a new software library for Euclidean Wasserstein-2 optimal transport.

DualAfford: Learning Collaborative Visual Affordance for Dual-gripper Manipulation

- Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, Hao Dong
- abstract@open-review(Poster): 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. Video demonstration can be found at https://sites.google.com/view/dualafford.

Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching

- Shengchao Liu, Hongyu Guo, Jian Tang
- abstract@open-review(Poster): Pretraining molecular representations is critical in a variety of applications for drug and material discovery due to the limited number of labeled molecules, yet most existing work focuses on pretraining on 2D molecular graphs. The power of pretraining on 3D geometric structures, however, has been less explored. This is owning 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 a 3D coordinate denoising pretraining framework to model such an energy landscape. Leveraging an SE(3)-invariant score matching method, we propose GeoSSL 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
- abstract@open-review(Poster): 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.

The Augmented Image Prior: Distilling 1000 Classes by Extrapolating from a Single Image

- Yuki M Asano, Aaqib Saeed
- abstract@open-review(Poster): 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 lucidpanda neurons" in networks that have never even seen one.

Delving into Semantic Scale Imbalance

- Yanbiao Ma, Licheng Jiao, Fang Liu, Yuxin Li, Shuyuan Yang, Xu Liu
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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.

Contextual Image Masking Modeling via Synergized Contrasting without View Augmentation for Faster and Better Visual Pretraining

- Shaofeng Zhang, Feng Zhu, Rui Zhao, Junchi Yan
- abstract@open-review(Poster): 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
- abstract@open-review(Poster): 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 (3+1)D Geometry-Sequence Modeling in Proteins

- Hehe Fan, Zhangyang Wang, Yi Yang, Mohan Kankanhalli
- abstract@open-review(Poster): 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 (3+1)D structure 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. Our code will be publicly available.